DOCUMENT RESUME

ED 193 247 TM 800 561

AUTHOR	Brown, Frederick G.
TITLE	Guidelines for Test Use: A Commentary on the Standards for Educational and Psychological Tests.
INSTITUTION	National Council on Measurement in Education, Washington, D.C.
PUB DATE	80
NOTE	132p.
AVAILABLE FROM	National Council on Measurement in Education, 1230 17th St. N.W., Washington, DC 20036 ($7.00).
EDRS PRICE	MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS	Codes of Ethics; *Educational Testing; *Guidelines; *Psychological Testing; Scoring; *Standardized Tests; *Standards; Test Bias; Testing; Testing Problems; Test Interpretation; Test Reliability; Test Selection; Test Validity
IDENTIFIERS	American Educational Research Association; American Psychological Association; National Council on Measurement in Education; *Standards for Educational and Psychological Tests

ABSTRACT

This monograph serves as a nontechnical guide to teachers, counselors, school psychologists, administrators, parents, and others who use the 1974 Standards for Educational and Psychological Tests. The Standards are a concise summary of standards and guidelines for the development, evaluation, interpretation, use, and reporting of information on psychological and educational tests and assessment methods. These guidelines focus on standardized and published assessment methods and, like the Standards, do not cover teacher-made tests. An introductory chapter identifies some of the basic considerations in educational and psychological assessment and provides a framework for the material discussed in following chapters. Chapter 2 considers the qualifications of test users, the process of selecting a test, and the standards relating to administration and scoring of the instrument. Chapter 3 focuses on the psychometric properties of the instrument, its reliability and validity. Chapter 4 discusses interpretation of test scores, and Chapter 5 covers several issues in test use, including test fairness and ethical issues. (RI)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED193247

OCT 14 1980

Guidelines for Test Use:

A commentary on the

Standards for Educational and Psychological Tests

U S DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TM800561

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Wm. Russell
NCME

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Frederick G. Brown

National Council on Measurement in Education

Guidelines for Test Use:

A commentary on the

Standards for Educational and Psychological Tests

Frederick G. Brown

Iowa State University

Copyright 1980
by the
National Council on Measurement in Education
All rights reserved

CONTENTS

Preface 5

I. Introduction 7

II. Test Selection and Administration 20

III. Reliability and Validity 42

IV. Interpreting Test Scores 71

V. Problems and Issues in Test Use 94

Glossary 116

References 126

Index 130

PREFACE

The primary purpose of this monograph is to serve as a user's guide and commentary on the Standards for educational and psychological tests. The Standards was developed by a joint committee of the National Council on Measurement in Education, the American Educational Research Association, and the American Psychological Association in response to the need for a concise summary of standards and guidelines for the development, evaluation, interpretation, use, and reporting of information on psychological and educational tests and assessment methods. The first edition, titled Technical recommendations for psychological tests and diagnostic techniques, was published in 1954. The current revision, published in 1974, is the third.

Although the Standards was designed to apply to all types of assessment techniques, we will focus on those methods used in educational settings. As such, this booklet will be of most interest to teachers, counselors, school psychologists, administrators, parents, and others (perhaps even students) concerned with educational measurement. Our goal is to present the information contained in the Standards in a nontechnical manner which will be understandable by persons who have not had extensive training in educational measurement, yet who use tests and other assessment methods in the educational process.

Several limitations of the presentation should be noted. First, the discussion focuses on standardized and published assessment methods; like the Standards, it was not designed to cover teacher-built tests. Second, reading this booklet should not be considered to be a substitute for formal courses in educational and psychological measurement or for supervised training in the administration and interpretation of tests. Third, the standards described should be considered as useful guidelines, not commandments. In the practical application of assessment methods, considerations other than technical adequacy (e.g., financial and situational constraints) play an important role. Nevertheless, the standards should be viewed as goals to work toward.

In writing this booklet I have been aided by many people. Particular thanks go to Nancy Cole, Jack Menne, and Jason Millman who reviewed drafts of the manuscript and made many valuable comments and suggestions; to Mrs. Gwen Ethington, who typed the manuscript; and to NCME, which provided the financial support for the project.

F. G. Brown
Iowa State University
Ames, Iowa

I am extremely pleased to write a brief preface for this booklet authored by my friend and colleague Dr. Fred Brown.

The mission or purposes of NCME are stated in the by-laws as follows:

(a) Encouragement of scholarly efforts to:

advance the science of measurement in the field of education;

improve measurement instruments and procedures for their administration, scoring, interpretation, and use; and

improve applications of measurement in assessments of individuals and evaluations of educational programs.

(b) Dissemination of knowledge about:

theory, techniques, and instrumentation available for measurement of educationally relevant human, institutional, and social characteristics;

procedures appropriate to the interpretation and use of such techniques and instruments; and

applications of educational measurement in individual and group evaluation studies.

NCME members have generally felt the Journal of Educational Measurement and NCME's annual meeting program meet the "scholarly effort" part of the mission. However, there has been constant concern about the "dissemination" aspect of the mission, particularly with respect to the "interpretation and use" of educational measurement techniques.

NCME is proud to present this clear and articulate monograph and hopes that it will serve to help the many teachers and administrators who make decisions concerning the interpretation and use of educational measurement techniques.

Jack Merrill, Ph.D.
Past President, NCME, 1978-79
Iowa State University
Ames, Iowa

I. INTRODUCTION

There are many situations where it is of interest or value to have information on the characteristics of individuals - their abilities, aptitudes, knowledge and skills, interests, values and attitudes, or personality characteristics. Sometimes individuals are interested in obtaining more information about their own abilities and characteristics. For example, a high school or college student may take ability and interest tests to help decide what curriculum or occupation to enter. Or an older adult, planning to re-enter college after being out of school for a number of years, may want to assess her academic skills and study habits to better prepare for the rigors of college work.

In other instances, one person may assess the characteristics of another individual. A common example is when teachers administer achievement tests to their students and use the results as a basis for assigning grades or planning instruction. Other examples come readily to mind: tests given for admission to graduate or professional schools; employment tests used when hiring new employees or deciding whom to promote; personality measures administered to determine what type of therapy or counseling to institute. In each case, the common core is that information derived from a test is used (by another person) to make a decision about the test taker.

Tests and other assessment methods may also be used for certification purposes. In these situations we are interested in whether a person has the minimum qualifications necessary to engage in some activity or practice an occupation. Common examples are tests (both written and on-the-road) taken to obtain a driver's license and examinations required for admission to the practice of a profession, - such as medical licensing and bar examinations.

In some cases the focus of testing may be on evaluating a program or institution rather than on individuals. For example, scores on achievement tests might be used to determine the relative effectiveness of several alternative instructional methods. More globally, information on students' achievements, skills, and accomplishments might be used as

part of an evaluation of the effectiveness of a school or curriculum. Note that, in these examples, even though we measure the characteristics of individuals, our goal is to evaluate the program, not the individuals enrolled in the program.

Purpose

In the broadest view, the purpose of this booklet is to promote the responsible use of tests and other assessment devices[1] in educational settings. Our discussion will be directed to persons who use tests and test results - e.g., teachers, counselors, personnel administrators, admissions officers - and people whose lives are influenced by tests - e.g., students, parents, teachers, job applicants, and employees. All these people need to be aware of the essential features and desirable characteristics of assessment devices, but do not need (or have) the technical sophistication of the professional psychologists who develop, apply, and/or do research on measuring instruments.

More specifically, our purpose is to present the material in the Standards for educational and psychological tests[2] in a less technical, though still rigorous form; in a manner that will be of more use to test users, particularly those who are not highly trained in psychometrics. The reader should have a copy of the Standards available when reading this booklet, as we will make continued references to it. We will also refer to texts on educational and psychological measurement and to journal articles, when appropriate. But our purpose is not to make this booklet a substitute for these sources, for courses in educational and psychological testing, or for supervised practice in test use. Rather we will try to present the essential considerations in selecting, evaluating, and using tests in a manner that is both understandable by test users and which will promote responsible use of tests.

We will attempt to accomplish this goal by discussing guidelines for the construction, evaluation, and use of tests. Some of these guidelines will involve the technical characteristics of tests - their reliability, validity, and normative data. Others will be oriented to the responsible use of tests; considerations involved in selecting, administering, and using tests and in interpreting test scores.

Such guidelines are necessary because: "Almost any test can be useful for some function and in some situations, but even the best test can have damaging consequences if used inappropriately" (p. 7).[3] That is, there are appropriate and inappropriate uses for any test. Our goal, therefore, is to help you develop the ability to differentiate between

appropriate and inappropriate uses. The placement of responsibility on you, the test user, is deliberate. Although test publishers have an obligation to provide information to users regarding the development and evaluation of their products, in the end there is no escaping the fact that "... the primary responsibility for the improvement of testing continues to rest on the shoulders of test users" (p. 7). Even a good test, if used improperly, may have a harmful impact.

Uses of Tests

Tests are methods of collecting information about an individual or a group of individuals. They are means rather than ends in themselves; thus, they can be evaluated only in the context of their stated aims or goals. Therefore, before considering the technical characteristics of tests and how they are used, we must first consider the broader goals of educational and psychological assessment.

As implied in the opening paragraphs, uses of tests can be divided into two broad classes: those concerned with providing a description of an individual's characteristics and those uses involving decision-making (Cronbach, 1971). Although sharing many features in common, there are several basic differences in these two approaches.

Descriptive uses. When using tests for descriptive purposes, we are not interested in collecting information to answer some immediate, specific question or to make a particular decision of consequence in a person's life. Rather we are amassing a data base that describes an individual; one that can be used many times for many purposes. Unless we are talking about self-administered and self-interpreted tests (a very rare commodity), the description is usually provided, at least in part, by the test user.[4] Here the goal of testing is to "... help the test user to understand, analyze, or help that individual" (p. 56). Examples are teachers using test results to plan instruction programs for their students, counselors using tests to help their clients make vocational choices, and clinical psychologists using test results to better understand their clients so as to facilitate communication within the therapeutic process.

Certain features of the descriptive use of test results deserve special mention. First, such uses are most appropriate and useful when there is a continuing relationship between the parties involved. For example, a teacher can use information from ability and achievement tests in a variety of contacts with her pupils, not just at one point in time. Or the therapist can use information from personality tests

throughout the course of therapy, not just at one juncture. The fact that there will be continuing contacts leads to a second point: The information may be modified in light of further data or used in other contexts, even ones unforeseen when the data were first collected. This latter fact, that the data can be continually reevaluated, might lead to the erroneous conclusion that tests used for descriptive purposes need not meet the technical standards demanded of tests used to make specific decisions. Such is not the case. Just because we are not faced with an immediate decision (e.g., whether to hire a particular person) we cannot infer that no important decisions will be made by individuals as a result of their consideration of the test data. Rather the decisions made are likely to be more global, long-range and, perhaps, ultimately more important to their lives. Thus, as good, if not better, information is needed.

One further question must be asked: How do we know what constitutes a good description? This is a validity question. While this question will be addressed in more detail in Chapter 3, we can provide a general answer at this point: "To validate a description is to answer the question: When persons are described in this way on the basis of these data, how much confidence can be placed in each of the implications of the description?" (Cronbach, 1971, p. 448). Note that we do not validate the description per se, rather we validate the implications made from the description. These implications will often go beyond the quantitative data (i.e., the test scores) as they will include the meaning the test user derives from a syntheses of the data.

Decision-making uses. Decision-making uses of tests are more direct and obvious to most persons. Here we have a choice between two or more alternative courses of action and the test data are used at least as a partial basis for making a decision. Some of these decisions are selection decisions. Should Yale admit Kirk Brown and/or Mark Anderson? Should the Acme Screw Company hire Paul Nutt as a management trainee? Others are placement decisions. Would Janice James do better as a production or a sales engineer with 4M Company? Should Mary Moppett enroll in Chemistry 11 or Chemistry 14? Should Zig Freud be treated with client-centered or behavioral therapy to cure his test anxiety? Other situations are harder to classify, yet clearly involve decisions. Should Ike Newton receive an A or B in calculus? Should Nancy Drew apply to Princeton or South Siwash, or to both schools?

In each case, there is a decision to be made, usually one that must be made relatively immediately and which will have consequences for the individual's life. The purpose of

considering the test data is to help make better decisions. Stated somewhat differently, "Testing is intended to reduce the number of incorrect predictions and hence the number of decisions that will be regretted later" (Cronbach, 1971, p. 448).

Evaluating the effectiveness of tests as decision-making aids is relatively straightforward. The question is one of obtaining an index of the effectiveness of the decision-making process. However, two further considerations must be kept in mind. First, we are evaluating the prediction made from the test, not necessarily the test per se. The prediction may be influenced by factors other than the test data (e.g., other variables, the test interpreter's biases), thus may go beyond the quantitative test scores. Second, predictions and decisions cannot be considered in a vacuum. The appropriate question is not how effective is this instrument as a decision-making aid? Rather it is: Is this instrument a more effective decision-making aid than other available methods?

Although descriptive and decision-making uses can be separated, it should be noted that the same data are often used for both purposes. Consider a student in vocational counseling who takes ability and interest tests as an aid in the continuing process of deciding upon career plans (a descriptive use). This student may also use the same information to decide whether to enroll in calculus next term or to apply to a particular college (decision-making uses). The important point to remember is that the evaluation of the effectiveness of each of these uses may require different sorts of evidence.

Specific goals and objectives. The descriptive and decision-making views represent two broad ways of looking at testing. When developing, evaluating, or using tests we have to consider even more specific goals and objectives. Consider the case of an arithmetic test designed for students in upper elementary grades. As the first step in developing the test we might decide that we want to measure arithmetic abilities and skills. But we have to go further. We must ask: What sorts of skills? For example, are we interested in computational skills, understanding of concepts, ability to apply skills, problem solving, or some combination of these? We must also define the range of content to be included: Will items use only whole numbers or will they include fractions and decimals? Do we include skills stressed in the "new math" or do we define the domain in more traditional terms? Do we want the test to indicate students' overall level of performance or do we want to diagnose strengths and weaknesses in particular subareas? And so on. The point is that there are many possible ways to define the goals and purposes of an instrument. Often all of them cannot be met by one instrument; thus some

goals are emphasized at the expense of others. Nevertheless, once defined, these goals provide a basis for the evaluation of the instrument.

Test publishers have an obligation to provide information regarding their goals to test users.

> B2. The test manual should state explicitly the purposes and applications for which the test is recommended. (p. 14)
>
> B3. The test manual should describe clearly the psychological, educational, or other reasoning underlying the test and nature of the characteristic it is intended to measure. (p. 15)

In other words, the test publisher must communicate to users what characteristic(s) the test is designed to measure and the ways in which the test can be applied.

Similarly, when considering various available tests, test users must consider their particular goals and evaluate the fit between their goals and those of the test publisher. If the two sets of goals are consistent, the test can be tried on a tentative basis. If the goals are not consistent, the test user is advised to look elsewhere, for the test will likely be of little help.

Characteristics of a Good Measuring Instrument

In our discussion we have used several terms interchangeably: test, assessment device, measuring instrument. At this point it would be well to clarify these definitions.

As in the Standards, the recommendations and guidelines in this booklet are designed to apply to any assessment procedure; that is, "to any systematic basis for making inferences about characteristics of people" (p. 2). This definition includes standardized aptitude and achievement tests, interest and personality inventories, attitude scales, and projective methods. It also includes interviews, biographical inventories, self-rating forms, checklists, and other clinical techniques. It could also encompass such ad hoc measuring devices as the typical classroom examination, although the Standards is not directed to this category of tests.

Tests. The most widely used assessment procedure, the test, can be viewed as a particular type of assessment device. The Standards defines a test as:

> ... a set of tasks or questions intended to elicit particular types of behavior when presented under standardized conditions and to yield scores that will have desirable psychometric properties such as high reliability and high validity. (p. 2)

Alternative definitions include:

> A test is a systematic procedure for observing a person's behavior and describing it with the aid of a numerical scale or a category-system. (Cronbach, 1970, p. 26)

> A psychological test is essentially an objective and standardized measure of a sample of behavior. (Anastasi, 1976, p. 23)

> A test (is) a systematic procedure for measuring a sample of behavior. (Brown, 1976, p. 7)

Other definitions of a test could be offered, but they would not differ significantly from those above. Thus, we see that the definition of a test is essentially the same as the (more general) definition of an assessment device.

To better understand what this definition implies, we can look at the component parts of the definition. Four points stand out. First, a test is a systematic procedure in that certain conditions are the same (or at least comparable) for all persons. In particular, all persons are administered the same or comparable items, under similar conditions, and responses are scored using a prescribed set of rules. Standardization of conditions is used for two reasons: a) to ensure that extraneous sources of variability do not influence test scores, and b) so that scores of different people can be compared to each other or to some performance standard. As an obvious example, consider a case where two people were administered different questions and/or had different time limits to complete the test. Because the conditions were not the same for each person, different factors could influence their performance; thus their scores cannot be directly compared.

Second, a test is a "sample" of behavior. What this means is that we seldom ask all possible questions, just some subsample of all possible questions. This sampling raises two questions: How do we know that the sample of questions used on the test adequately represents the broader domain of interest? And, would the test taker obtain the same score on another sample of items from the same domain? The former is a question of validity, the latter one of reliability.

Third, a test measures "behavior". In the strictest sense, this behavior is the marks, words, or symbols the test taker places on the answer sheet; or, on situational and performance measures, the verbal and motor responses made. From these responses, and our knowledge of what the test measures, we infer something about the characteristics of the person making the responses. Thus, our measurement is always indirect and inferential. This indirect nature of the measurement process highlights a very important point: "A test score describes but it does not explain a level of performance" (p. 2; underlining added). That is, a test measures the test taker's behavior, abilities, or characteristics at a given point in time. To understand why a person obtained the score he did usually requires further information.

Finally, a test is a measuring instrument. Measurement is a complex concept which, in its simplest form, refers to the assignment of numerical values, along some continuum, to performance in accordance with specified rules.[5] This means that, using certain prescribed rules, the test taker's responses will be translated into a numerical value. To fully understand the meaning of the numbers assigned requires knowing both the nature of the continuum on which the scores are expressed (i.e., the nature of the characteristic being measured) and the operations used in transforming the responses to a numerical value (i.e., the scoring and scaling procedures used).

Essential characteristics. Given this definition of a test we can now ask: What are the essential characteristics of a good test? These characteristics fall in four categories: test construction, test administration, psychometric properties, and interpretive procedures. In this section we will give an overview; in later chapters we will discuss each aspect in more detail.

The essential elements in the test construction phase are for the test developer to clearly specify what the test measures and to explain how the item selection and test development process ensure that the test will measure what it is designed to measure. Given these specifications, the publisher may also suggest appropriate uses for the instrument.

The major concern regarding test administration is to develop procedures that will make testing conditions comparable for all test takers. The test publisher must develop and specify standard procedures (e.g., directions, time limits, response modes) for administering the test and a set of rules for scoring the responses. Furthermore, these procedures should be such that the prospective test user can follow them

without inordinate difficulty. These rules should be clearly communicated to the user by the test manual.

Psychometric adequacy revolves around two questions. The first is whether the test measures (whatever it measures) consistently. That is, if the test taker took a different sample of items or took the same items at a different time, would he obtain (approximately) the same score? This is the question of reliability. The other psychometric consideration is whether the test measures what it is intended to measure. This reduces to determining what the test does, in fact, measure. This is the validity question.

Finally, test scores must be interpreted - someone must make sense out of the test taker's responses. Besides expressing performance on some meaningful scale, there must be a basis for making comparisons, either between persons or against some standard. The former comparison requires normative data (and usually transformation of scores to a more meaningful scale, such as percentile ranks); the latter requires some rational or empirical basis for setting a performance standard. And, as test results will be useless unless their meaning can be communicated to the persons concerned with the scores, there must be procedures available for making the meaning of the scores intelligible and to prevent, in so far as is possible, misinterpretation of scores.

Provision of most of the information described above is the responsibility of the test publisher or developer. This information should be provided in a manual in a form that the intended user can understand. Most of the remainder of this booklet will be concerned with specifying guidelines the potential test user can use to determine if the necessary information for intelligent, appropriate, and ethical test use is available. Before doing this, however, let us briefly consider two other points.

Which Test or Tests?

In our discussion we have made two implicit assumptions: that an already developed instrument met our purposes, and that the necessary data could be obtained from one instrument. Neither of these assumptions is necessarily correct. Let us look at each assumption briefly.

The availability of tests. As pointed out previously, the first step in developing an assessment procedure is to specify one's goals. Or, as stated in the Standards:

H1. The choice or development of tests, test batteries,

> or other assessment procedures should be based on clearly formulated goals and hypotheses. (p. 60)

Suppose, as a test user, that after you have formulated your goals, you check through the various sources of information about available tests and find the cupboard is bare; there are no instruments that appear to fit your needs. What do you do now?

In this situation you have three choices. One is to forget about the problem at hand and/or adopt some arbitrary method of making a decision. This procedure has certain obvious limitations. A second possibility is to rely on some informal, nonstandardized method of assessment. For example, employers often rely on judgments formed during interviews with job applicants rather than on more standardized assessments of candidates' skills and abilities.[6] This procedure also has severe limitations which, if not clear now, will be by the time you finish reading this booklet. This leaves a third possibility, to develop your own assessment device. This option generally requires much time and effort; however, in many situations the only viable alternative is to construct your own assessment device.[7]

Need for multiple assessments. Rarely, in practice, will use of a single assessment device suffice. As the Standards states:

> For most purposes, the evaluation of a person requires description that is both broad and precise; a single assessment or assessment procedure rarely provides all relevant facets of a description.
>
> Decisions about individuals should ordinarily be based on assessment of more than one dimension; when feasible, all major dimensions related to the outcome of the decision should be assessed and validated. (pp. 61-62)

Or expressed as a guideline:

> H2. A test user should consider more than one variable for assessment and the assessment of any given variable by more than one method. (p. 61)

In other words, because human behavior is complex, our assessment procedures generally will consider several dimensions simultaneously.

The most obvious implication is the one stated above: The test user will often need to consider several dimensions

or measures simultaneously. A second implication is also contained in H2: The test user may, in some situations, want to measure a characteristic by more than one method. This point is often overlooped. Different methods of measuring a trait (characteristic, dimension) may not produce the same information. For example, will students' achievement be the same when assessed by multiple-choice and essay items? Or, will a person display the same level of dominant behavior in a social interaction as he proclaims on a self-report inventory? Thus, we must always consider how we measure the characteristic as well as what characteristic is being assessed.

A third implication is also germane to our discussion. Often data from several assessment sources are combined to form a global description or to predict a particular outcome. Sometimes data from several different measures are used; for example, scores on college admissions tests and high school grades may be combined to predict college grades. In other instances, data on several characteristics may be obtained from the same instrument; for example, the various subscales of an intelligence test or the various tests comprising an achievement test battery. Whenever data from several assessment sources are combined, the validity of the overall assessment will depend on several factors, both ones relating to the individual measures and ones related to the method used to combine the data. The methods used to combine data are many, varied, usually complex, and beyond the scope of this discussion. Thus, we will concentrate on guidelines for determining whether single measures are valid and useful; we will leave to other sources the questions of how to best combine measures.[8]

Summary and Plan

In this chapter we have identified some of the basic considerations in educational and psychological assessment, thus providing a framework for the material to be discussed in the following chapters. We defined the term "test", described the essential characteristics of a good measuring instrument, and differentiated between descriptive and decision-making uses of tests. Throughout the discussion we emphasized that any measuring instrument must be evaluated with reference to specified goals and objectives, both those specified by the test developer and those specified by the test user. The latter will generally be more important as they reflect the particular situation in which the test will be used.

In the following chapters we will discuss the various aspects of the assessment process in greater detail. In Chapter 2 we will consider the qualifications of test users, the process of selecting a test, and the standards relating to

administration and scoring of the instrument. Chapter 3 will focus on the psychometric properties of the instrument, its reliability and validity. Chapter 4 will discuss considerations relating to the interpretation of test scores. Finally, in Chapter 5, we will consider several issues in test use, including test fairness and ethical issues. Our goal will be to better prepare you, the reader, to make more informed judgments about the problems and issues concerning the selection, evaluation, interpretation, and use of tests and other assessment methods in educational settings.

Notes

1. We will use the term "test" to refer to any assessment device. The reader should remember, however, that the guidelines discussed in the Standards apply to all types of assessment devices, both those commonly called tests and those that go under other names (e.g., inventories, checklists).

2. The Standards for educational and psychological tests was written by a joint committee of the National Council on Measurement in Education, the American Educational Research Association, and the American Psychological Association. The standards were designed to apply to all assessment procedures but were written with specific reference to standardized tests. Included are standards relating to test development, research studies on reliability and validity, interpretation and use of test scores, and information to be provided in test manuals.

3. Quotations from the Standards will be referenced with only a page number, e.g., (p. 56). A list of references can be found at the end of the booklet.

4. The term "test user" refers to "one who chooses tests, interprets scores, or makes decisions based on test scores" (p. 1). A glossary of terms can be found at the end of this booklet.

5. See Jones (1971) or Kerlinger (1973, Chap. 25) for readable discussions of the concept of measurement.

6. Informal assessment methods are often used even when standardized methods are available. Our point here, however, is that these methods can be used when no standardized measures are available.

7. For a discussion of constructing achievement and ability test see, e.g., Bloom et al. (1971), Brown (1976), Ebel (1972), Mehrens & Lehmann (1978), Popham (1973, 1978),

Notes (Chapter 1, Cont'd)

Thorndike (1971a), or Thorndike & Hagen (1977). For a discussion of personality inventories and attitude scales see, e.g., Anastasi (1976), Brown (1976), Dawes (1972), Edwards (1970), Nunnally (1978), or Wiggins (1973).

8. See, e.g., Brown (1976), Cooley (1971), Kerlinger & Pedhazur (1973), or Nunnally (1978).

II. TEST SELECTION AND ADMINISTRATION

The effective and ethical use of tests depends on three factors: (1) the technical quality of the assessment instrument, (2) how and where the test is used, and (3) the test user. Weakness in any one of these areas can interfere with optimal use of the test. Consider use of a vocational interest inventory. There are numerous interest inventories on the market and they vary widely in quality. Selecting one of the poorer inventories will result in less useful information than using one of the better inventories. But suppose that we decide to use one of the better inventories. Our next question is, how will it be used? As scores on interest inventories are, at least to some extent, susceptible to faking, the test may provide useful information in situations where examinees will give honest responses, such as in vocational counseling. However, in situations where test takers may have reason to bias their responses (for example, if the test were used as a selection device) the results may be more open to question and less valid. Even if a good test is selected and used in an appropriate situation, test users who are not thoroughly familiar with the test and the literature on vocational interest development may draw improper inferences from the test results. Thus, we can see that any one of the three conditions may result in inappropriate use of the test.

In this chapter we will discuss three factors relating to effective use of assessment methods: the qualifications of test users, how to select a particular assessment method, and the administration and scoring of tests. In following chapters we will discuss the psychometric properties of measuring instruments, the interpretation of test scores, and problems and issues in assessment.

Qualifications of Test Users

We have all taken many tests - classroom exams, standardized achievement and aptitude tests, tests to obtain a driver's license, tests to obtain a job or become certified in our profession, and so on. In most of these situations the testing procedures probably seemed so simple and straightforward that anyone with even a modicum of intelligence could

administer and interpret the test. But in testing, as in other areas of life, things are not always as simple as they seem. Thus, we need to look at the qualifications needed by test users.

First we must define more clearly who we mean by a test user. In the Standards the term "test user" refers to "one who chooses tests, interprets scores, or makes decisions based on test scores" (p. 1). This may or may not be the person who administers and scores the test; in many situations, administration and scoring is done by a trained clerk or semi-professional. The test users we are concerned with are people who interpret and make decisions based on test results. This, of course, requires training and knowledge beyond routine clerical skills. Anastasi makes this point very clearly:

> The need for a qualified examiner is evident in each of the three major aspects of the testing situation - selection of the test, administration and scoring, and interpretation of scores. Tests cannot be chosen like lawn mowers, from a mail-order catalogue. They cannot be evaluated by name, author, or other easy marks of identification. To be sure, it requires no psychological training to consider such factors as cost, bulkiness and ease of transporting test materials, testing time required, and ease and rapidity of scoring. ... For the test to serve its function, however, an evaluation of its technical merits in terms of such characteristics as validity, reliability, difficulty level, and norms is essential. Only in such a way can the test user determine the appropriateness of any test for his particular purpose and its suitability for the type of persons with whom he plans to use it. (Anastasi, 1976, p. 31)

Though not considered in this passage, similar comments could be made regarding the interpretation and use of test scores.

Responsibilities of the test publisher. Determining who is a qualified test user is the responsibility of two parties, the test publisher and the test user. The test publisher's responsibility is to inform potential users of the qualifications needed to use the test.

> B4. The test manual should identify any special qualifications required to administer the test and to interpret it properly. (p. 15)

These qualifications are usually thought of in terms of special training (e.g., courses or degrees) needed to achieve competence, but it is possible, in some cases, to specify

qualifications in competency-based terms (Mayo, 1964, 1967).

Being a qualified test user is not an all-or-none thing. Because tests are used for various purposes, many users will have some, but not all, the desired qualifications. The manual should describe the qualifications needed for each use.

> B4.2. Where a test is recommended for a variety of purposes or types of inference, the manual should indicate the amount of training required for each use. (p. 16)

If the potential user has the qualifications to use a particular type of instrument, but lacks experience with a specific test, the manual should point out useful sources of information.

> B4.3. The manual should draw the user's attention to references with which he should become familiar before attempting to interpret the test results. (p. 16)

The test manual should also indicate the most frequent types of errors in interpretation and/or use of the test and suggest training or other means to eliminate the errors (B1.4).

In an earlier edition of the Standards (1966), a distinction was made between three levels of tests:

> Level A. Tests or aids that can be adequately administered, scored, and interpreted with the aid of the manual and a general orientation to the kind of institution or organization in which one is working (e.g., achievement or proficiency tests).
>
> Level B. Tests or aids that require some technical knowledge of test construction and use, and of supporting psychological and educational fields such as statistics, individual differences, psychology of adjustment, personnel psychology, and guidance (e.g., aptitude tests, adjustment inventories applicable to normal populations).
>
> Level C. Tests and aids that require substantial understanding of testing and supporting fields, together with supervised experience in the use of these devices (e.g., projective tests, individual mental tests). (pp. 10-11)

Although these levels are not completely explicit or independent, they do suggest the levels of training and competency required to interpret and use various classes of tests.

The thrust of all these recommendations is clear - the

test publisher is responsible for telling potential users what qualifications they should possess in order to interpret and use the test.

Test user responsibilities. While test publishers are responsible for informing potential users of the qualifications necessary for proper use of a test, users must ultimately decide whether or not they possess these qualifications. Thus, the Standards states:

> G1.1. A test user should know his own qualifications and how well they match the qualifications required for the uses of specific tests. (p. 58)

This, of course, is much easier to state than to accomplish.

What are these desired qualifications? Certainly the user should have some basic knowledge of the principles of psychological testing:

> G1. A test user should have a general knowledge of measurement principles and of the limitations of test interpretations. (p. 58)

Or, more specifically:

> The required level of knowledge will vary with the complexity of the evaluations to be made and the responsibility of the user. At a minimum, the user must be knowledgeable about testing principles, understand the concept of measurement error, and be able to interpret an obtained test score. He should realize that there are alternative explanations for a given score and should have a pool of knowledge from which to evaluate some of the alternatives. (p. 58)

> G3.1.1. A test user should have sufficient technical knowledge to be prepared to evaluate claims made in a test manual. (p. 59)

This level of knowledge, for all but the least complex tests, would probably require an introductory course in educational and psychological testing.

Besides having the qualifications necessary to evaluate a test, users must also consider the social and psychological context in which they are operating.

> G2. A test user should know and understand the literature relevant to the tests he uses and the testing problems

with which he deals. (p. 58)

H3. In choosing an existing test, a test user should relate its history of research and development to his intended use of the instrument. (p. 63)

Several examples come readily to mind. One, used as an illustration in the Standards, is the (improper) use of intelligence tests to classify culturally disadvantaged and non-English speaking students as mentally retarded. Others include the use of unvalidated tests to label children as learning disabled and the use of tests measuring knowledge of educational principles as measures of the effectiveness of teachers.

If the test is to be used to make decisions about individuals, the test user has further responsibilities.

G3. One who has the responsibility for decisions about individuals or policies that are based on test results should have an understanding of psychological or educational measurement and of validation and other test research. (p. 59)

G3.1. The principal test users within an organization should make every effort to be sure that all those in the organization who are charged with responsibilities related to test use and interpretation (e.g., test administrators) have received training appropriate to those responsibilities. (p. 59)

G3.2. Anyone administering a test for decision-making purposes should be competent to administer that test or class of tests. If not qualified, he should seek the necessary training regardless of his educational attainments. (p. 59)

The implications of these principles are clear - all persons concerned with test use, at whatever stage or level of responsibility, should thoroughly understand what they are doing. If they do not have sufficient training or information to use the test properly, they should obtain it. Otherwise, they should not be using the test.

Test Selection

The first step in any assessment process is to define one's goals and purposes. An equally as important second step is to select the particular assessment method to use. As mentioned previously, the test user has several alternatives: to

select a published test or some combination of published tests, to develop a new test, or to rely on other methods of assessment. In this section we will be concerned only with selecting and evaluating published instruments.

Sources of information. By defining her purposes and considering the target population, the test user has taken the first step in narrowing the search. For example, the test user may want an American history test for junior high students, a reading readiness test appropriate for middle-class first grade children, or an academic ability test suitable for use in the upper elementary grades.

Given these requirements, how can one find out what tests are available? Usually, the first place to look will be in Tests in Print II (Buros, 1974). This is a classified, annotated listing of approximately 2500 published tests. While not evaluating any tests, Tests in Print II does provide an index to commercially published tests.

The user should next consult The Mental Measurements Yearbooks (MMY). The seventh edition was published in 1972 (Buros, 1972) and the eighth edition in 1978 (Buros, 1978). The Yearbooks include basic information about published tests (e.g., scores, time limits, cost, publisher), critical reviews of specific tests written by testing and subject matter experts, and bibliographic references to studies using each test. Not all tests are reviewed in each edition, but most new, revised, and widely-used tests are reviewed. References to reviews in previous editions of the MMY can be found both in the current edition and in Tests in Print II. Any potential test user should consult the MMY before selecting a test.[1]

Another useful source of critical reviews of tests is the publications of the Center for the Study of Evaluation at UCLA. Their monographs critically review tests in several areas: e.g., elementary school tests; tests for higher-order cognitive, affective, and interpersonal skills; preschool/kindergarten tests; criterion-referenced tests. Other sources of information include catalogs of the various test publishers, textbooks on psychological and educational testing, journals in the area of educational and psychological measurement, the ERIC bibliography (Backer, 1977), faculty members at colleges and universities, and directors of school testing programs.

These sources provide information on the availability and quality of published tests, thus helping the user to narrow the search to a small number of potentially useful instruments. Before adopting a particular test, the user should

review each test in detail. The best way to do this is to obtain specimen sets of the tests being considered from the test publishers. A specimen set (available at moderate cost) contains copies of the test and supporting materials such as technical and administrator's manuals, interpretive aids, and descriptions of auxilliary services. Careful study of these materials is essential in order to properly evaluate a test.

The test manual. The user's primary concern is with the nature, quality, and potential usefulness of the test. To evaluate these characteristics requires looking at the test itself and any supplemental materials (answer sheets, scoring keys, profiles, interpretive aids, etc.). But the user also needs to have information concerning the rationale, development, administration, interpretation, and technical adequacy of the test. This information should be in the test manual(s). Thus, provision of manuals for a test is a necessary part of the test developer's task.

> A1. When a test is published or otherwise made available for operational use, it should be accompanied by a manual (or other published or readily available information) that makes every reasonable effort to follow the recommendations of these standards and, in particular, to provide the information required to substantiate any claims that have been made for its use. (pp. 9-10)

Thus, the test manual is a prime source, but by no means the only source of information about a test.

The manual should provide the user with the information needed for accurate and appropriate use of the test. Again, to quote the Standards:

> For each test there should be a test manual, perhaps with supplements, to provide enough information for a qualified user to make sound judgments regarding the usefulness and interpretation of test scores. (p. 5)

> A test user needs information to help him use the test in standard ways and to evaluate a test relative to others he might select for a given purpose. The information that he needs to select a test or to use it must come, at least in part, from the test developer. (p. 9)

In addition to the basic information about the test, the test manual should also indicate what sorts of information are not included in the manual and the importance of such material (A1.1), give references to other sources of information about the test (A1.2), and provide information about any short or

abridged forms of the test (A3.2.1, A3.2.2).

In addition to providing information about the published forms of the test, the manual should describe the development of the test, particularly the rationale and procedures for selecting the items or tasks included on the test.

> A2. A test manual should describe fully the development of the test: the rationale, specifications followed in writing items or selecting observations, and procedures and results of item analysis or other research. (p. 11)

It is generally useful to include sample items. If, for reasons of test security, actual items cannot be included, the manual should at least provide illustrative or sample items so the potential user can judge their nature and quality (A2.3). The manual also should include a non-technical description of the rationale for the items. An example of how this can be accomplished is found in the manuals for many achievement tests, which list the content and skills measured by each test item. For example, item 1 may measure the ability to apply the addition operation to two-digit numbers, item 2 the ability to apply decimals to a problem of making change, and so on. This information allows the user to better understand what the test is trying to measure. When the content domain is not as clearly specifiable as in previous examples, the manual should give the qualifications of the persons who wrote the items (A2.4).

One should also keep in mind that tests, like other products, may become dated. A test may become dated in terms of its content (e.g., a science test published ten years ago may not accurately reflect the science being taught in today's schools) or evidence may have been collected that casts doubt on the validity or usefulness of the test. Thus:

> A3. The test and its manual should be revised at appropriate intervals. The time for revision has arrived whenever changing conditions of use or new research data make any statements in the manual incorrect or misleading. (p. 12)

When revising a test and/or manual, the test publisher has an obligation to take the research evidence, be it positive or negative, and changing social and educational conditions into account (A3.1, A3.2). Test publishers should also provide whatever information is necessary to inform users as to what changes in the test, if any, have been made and provide data on the comparability of the old and revised forms. The user's obligation is to consider what effect the recency of the test

content, norms, and technical data will have on the quality of the test and its use.

When reading a test manual, the user should consider the implications left by the manual as well as its literal statements.

> A manual is to be judged not merely to its literal truthfulness, but by the impression it leaves with the reader. If the typical professional user for whom the manual is prepared is likely to obtain an inaccurate impression of the test from the manual, the manual is poorly written. The standards apply to the spirit and tone of the manual (or supplemental publication) as well as to its literal statements. (p. 5)

While most test developers and publishers are honest and ethical people, like others who have something to sell, they want to present their product (the test) in the best light. Test users should keep this fact in mind when reading test manuals.

In summary, the test publisher has an obligation to provide accurate information about the test to the potential user, information that will allow users to judge the adequacy of the test for their purposes. This obligation extends not only to the manuals, but also to any promotional material used to market the test (A1.2.2). Test users have the obligation to critically consider and evaluate all the materials relating to the test, then use their best professional judgment when deciding whether to adopt the test.

What to look for: evaluating the test. Once the search has been narrowed to several possible instruments, the potential user must look at each test critically and make a final choice. As we have so often reiterated, this choice must be made in light of the goals and constraints of the particular situation. While, in one sense, this entire booklet can be considered as a guide to selecting the best test and using it effectively, let us once again quote the most relevant general principles from the Standards:

> B2. The test manual should state explicitly the purposes and applications for which the test is recommended. (p. 14)

> B3. The test manual should describe clearly the psychological, educational, and other reasoning underlying the test and the nature of the characteristic it is intended to measure. (p. 15)

> A2. A test manual should describe fully the development of the test: the rationale, specifications followed in writing items or selecting observations, and procedures and results of item analysis or other research. (p. 11)
>
> B5. Evidence of validity and reliability, along with other relevant research data, should be presented in support of any claims being made. (p. 16)
>
> B1. The test, the manual, the record forms, and other accompanying material should help users make correct interpretations of the test results and should warn against common misuses. (p. 13)

Any test evaluation will be directed towards assuring that these conditions are met.

In any given situation, other factors may have to be considered when selecting an assessment method. Two are particularly worthy of note. The first is practical considerations, such as cost (more particularly, cost effectiveness) and time limitations. These constraints often operate independently of the technical quality of the test. That is, if a school has only so many dollars or so much time available for testing, use of a particular test may be ruled out. The important point, however, is that practical matters are always of lesser importance than the quality of the instrument; a poor test should never be used solely because of its practical advantages.

The other point is the presence or appearance of bias in the test or testing procedures.

> G4. Test users should seek to avoid bias in test selection, administration, and interpretation; they should try to avoid even the appearance of discriminatory practice. (p. 60)

As the Standards points out, this is a very difficult standard to apply, both because people disagree on the meaning of the term "bias" and because we do not fully understand the sources of bias and how they operate. Yet, we must always be sensitive to possible biasing effects. One practical suggestion would be, whenever possible, to have qualified representatives of any group that might find the best to be biased or objectionable review the test before it is adopted for use.

But how does one actually make a test evaluation? This is not the place to discuss this process in detail. Examples of questions to be asked when evaluating tests are given in

Anastasi (1976, Appendix C), Brown (1976, pp. 454-455), Payne (1974, pp. 344-346), Thorndike & Hagen (1977, pp. 107-110), and in other texts. The interested reader is referred to these sources.

Summary. The process of selecting and evaluating a test involves first defining one's objectives, then applying the standards to see whether the test is technically acceptable and fits your purpose. This is a rational, judgmental process which requires knowledge of testing and of the situation in which the test will be used. As various concerned persons may have different perspectives on the testing program, it is usually a good idea to have the test reviewed by several people, not just one. For example, in a school setting one should consult with the testing director, school psychologists and counselors, administrators, classroom teachers, and perhaps even parents and students, before making a final choice of an instrument.

When selecting a test, the choice should never be considered as final; rather it should be considered as the best choice for the moment. The test use and procedures must be constantly monitored and evaluated to ensure that the test is being used correctly and effectively.

> G5. Institutional test users should establish procedures for periodic internal review of test use. (p. 60)

This evaluation should consider not only the effectiveness of the particular instruments used but also the roles, functions, and values of the entire testing program.

Administering and Scoring the Test

The procedures for administering and scoring a test usually appear to be relatively straightforward and simple. And in most cases, they are - a conspicuous exception being individually administered tests, such as the Stanford-Binet and Wechsler Intelligence Scales. Certainly most educational tests can be administered and scored by anyone who is willing to follow directions and use common sense. This is not to imply that these tests can be interpreted and used routinely; only that they can be administered and scored by persons with a minimum of special training.

When administering and scoring a test, the goal is to obtain a sample of the person's behavior under standard conditions. Use of controlled (i.e., standard) conditions is necessary for two reasons. One is to minimize, or hold constant, the effects of extraneous variables (e.g., time limits,

guessing) which might influence the scores obtained. If these factors were not controlled, scores might be quite variable over occasions and, thus not be dependable. The second reason is so scores of different persons can be compared. Unless two (or more) persons take a test under the same conditions, we cannot directly compare their scores.[2] Thus, the emphasis on using systematic, controlled procedures and observations.

More broadly, care is needed when administering and scoring tests so as to obtain the most accurate data on which to base our inferences. Remember, our concern in assessment is not with the scores per sé, but with the inferences made from the scores. For example, we are interested in the scores Dan makes on a reading test only because these scores allow us to infer something about his reading skills and, thereby, plan further instruction for him. If these inferences are based on test scores that reflect, in part, irrelevant factors, our inferences will be less valid. Thus, the need for care in administering and scoring tests.

Administration: Directions. As a test is being developed, procedures are devised for administering the test. These procedures, in turn, must be communicated to the test user. This is done through the test manual, as a set of directions to be followed when administering the test. Unless there is some compelling reason for changing procedures (see below), the test administrator should follow the procedures exactly as they are stated in the manual. Otherwise, the standardization data will not apply.

These directions are addressed to two groups of people, test administrators and test takers. Consider first the test administrator.

> C1. The directions for administration should be presented in the test manual with sufficient clarity and emphasis so that the test user can duplicate, and will be encouraged to duplicate, the administrative conditions under which the norms and the data on reliability and validity were obtained. (p. 18)

That is, after studying the manual the test administrator should be able to duplicate the administrative procedures used in standardizing the test. How complete these directions must be will vary with the nature of the test and the population being tested. For example, directions for a test consisting of several separately-timed parts and/or containing unfamiliar types of items or response procedures will necessarily be more complex than those needed on a multiple-choice achievement test.

Two points that frequently cause confusion are worth noting. One, what about questions from test takers?

> C1.1.2. The directions to the test administrator should include guidance for dealing with questions from examinees. (p. 18)

While test publishers cannot be expected to anticipate all questions, they should at least provide guidance on acceptable responses to common questions, such as whether examinees should be instructed to guess or can be told the meaning of a word used in an item.

The second question is more complex: Under what conditions can, or should, testing conditions be modified to suit a particular individual or situation? The general answer is rarely. Yet there are circumstances when such modifications may be appropriate or desirable. For example, a teacher might want to extend the time limits on a test to see how a student would perform when not under time pressure; or questions might have to be read to a student with a visual handicap. Here, again, the test publisher should provide some guidance to the test user.

> C1.2. If expansion or elaboration of instructions described in the test manual is permitted, the conditions under which this may be done should be clearly stated either in the form of general rules or in terms of giving numerous examples, or both. (p. 18)

If such modifications are made, the test user must realize that the obtained scores cannot be directly compared to those of persons who took the test under the standard conditions.

Directions must also be provided for the examinees. Here, the primary concern is that test takers' scores represent their performance on the test tasks and are not influenced by a misunderstanding of the test procedures.

> C2. Instructions should prepare the examinee for the examination: Sample material, practice use of answer sheets or punch cards, sample questions, etc., should be provided. (p. 18)

The nature and extent of such preparation will vary with the experience and maturity of students and the complexity of the testing procedures. For example, directions for filling out electronically-scored answer sheets may be needed for young children first encountering such answer sheets but will probably be unnecessary (or minimal) for college students.

Certainly the directions for examinees should cover all common questions, such as time limits, where and how to respond, and whether to guess when unsure of the correct answer.

> C1.1.1. The directions should clearly point out such critical matters as instructions on guessing, time limits, and procedures for marking answer sheets. (p. 18)

The guiding principle in preparing examinees should be: Provide all necessary information on how to respond but none on what to respond.

Administration: The test user. The test publisher has an obligation to provide clear and complete directions for administering the test, directions that preserve the standardization (comparability) of conditions across situations. But this is only part of the story. Unless test users follow these directions, interpretation of obtained scores with reference to the standardization (normative) data will not be possible. Thus:

> I1. A test user is expected to follow carefully the standardized procedures described in the manual for administering the test. (p. 64)

> I1.1. A test user must fully understand the administrative procedures to be followed. (p. 64)

That is, it is the responsibility of test users to study the manual, thoroughly understand what is required of them and of the test takers, then apply the directions in the manner prescribed in the manual. Failure to do so may invalidate the results obtained.

Oftentimes the administration of a test will be delegated to someone other than the test user. In these cases, the user retains responsibility for the test administration and must ascertain that the administrators are qualified to perform their task. This may require training the test administrators and monitoring their performance to see that the desired procedures are being used.

The test user also has the responsibility for determining that all the testing materials are in usable conditions.

> I1.3. A test user should make periodic checks on material, equipment, and procedures to maintain standardization. (p. 64)

Examples include seeing that enough tests and answer sheets

are available, discarding or repairing test booklets that have been marked by previous users, checking the accuracy of stop watches or other timing devices, and determining whether any mechanical or physical apparatus is in working order. As noted above, this responsibility also extends to monitoring the performance of the test administrator (including oneself, if the test user is also the administrator). It is too easy, especially after administering a test a number of times, for sloppy or bad habits to develop.

Testing conditions. We have emphasized the need for the test administrator to follow the standard directions for testing. This is a necessary requirement. But there is an equally as important function of the test administrator - to establish the proper atmosphere or climate in the testing situation. As testing is usually viewed as an evaluative situation by test takers, and as many people have evaluation apprehension, it is necessary to establish optimal testing conditions.

> I2. The test administrator is responsible for establishing conditions, consistent with the principle of standardization, that enable each examinee to do his best. (p. 65)

The phrase "do his best" can be interpreted literally with maximal performance tests; with typical performance measures the phrase can be interpreted to mean "present an honest picture of himself."

The meaning of this standard is clearly explained by the comment following principle I2 in the Standards:

> In a negative sense, the goal of this standard is that conditions inhibiting maximum performance should be avoided. The principle can be followed in part simply by being sure that all materials - such as answer sheets, pencils, and erasers - are on hand and that precautions have been taken to avoid distractions. In a more positive sense, the administrator should be sure that the examinee understands the tasks involved in taking the test: what kinds of responses are to be made and on what answer sheets, the implications for test-taking strategy of erasures or multiple marking or guessing, and how to know whether the test has been completed.
>
> The tester should try to create a nonhostile environment; standardized procedures are impersonal, but the test administrator must avoid being either patronizing or unresponsive to the examinees, especially when the tester

> and the examinee differ in race, sex, or status. A testing situation contains elements that are nonrecurring and unique to the persons tested. Although these may have negligible effects on test reliability, they may include events perceived as denigrating or questioning of the worth of the individual. A complete catalog of such events is not possible or easily described. In general, however, the social amenities of respect, politeness, and due regard for extenuating circumstances are relevant guides for ensuring the dignity of persons. While it may not be demonstrated that abuse of these principles leads to poor test performance, such abuse is not likely to enhance performance.
>
> It is often difficult to maximize the motivation of the examinees. The attempt is important; a major source of error may arise when examinees do not like or trust the test, tester, or test situation, and therefore make no special effort to do well in it. (p. 65)

This comment stresses the need for developing rapport with the examinees, relieving test anxiety, developing and maintaining motivation, and avoiding bias. Certainly training of examiners should include discussion of these topics.[3]

The overall goal of test administration is to create conditions which minimize the effects of the tester and test conditions so that scores will accurately reflect the characteristics of the test takers:

> H4. In general a test user should try to choose or develop an assessment technique in which "tester-effect" is minimized, or in which reliability of assessment across testers can be assured. (p. 63)

The rationale is that the less the influence of the tester and testing conditions, the fairer the test will be and the more accurately the test will reflect the examinee's performance. Certainly, this goal is easier to obtain on certain types of tests; e.g., examiner effects will be more pronounced on individually administered intelligence tests than on group achievement tests. In all cases, however, using standardized procedures is the best way to minimize testing effects.

Scoring the test. The next step is to score the test. Here our goal is to obtain objective scoring. By "objective" scoring we mean use of procedures that ensure agreement between two or more competent (trained) scorers. This does not mean there can be only one correct answer for each item or that anyone can score the test. What is necessary is that

procedures be established such that trained scorers will agree on the scores assigned. Ideally, there would be perfect (100%) agreement between scorers; in practice this is not always possible. For example, on free-response items (e.g., essay questions, some items on individual intelligence tests), scorer judgment may be needed when assigning point values to responses. But our goal should always be to work towards perfect agreement among scorers.

Objective scoring entails three requirements. The first is immediate and unambiguous recording of responses. On alternative-choice items this will be the mark on the answer sheet; other types of items may require verbatim recording of the examinee's responses. The second requirement is a list of standard or correct responses, a scoring key. This provided by the test developer and can vary from a list of correct responses or a scoring template (on alternative-choice tests) to a detailed manual, such as those provided with individual intelligence tests. The third requirement is specified procedures for comparing the examinees' responses to the key and obtaining a score. These directions should consider such matters as what weight to assign to each item or response, how to assign credit for incomplete or partially correct answers, and corrections for guessing.

The directions and procedures for scoring the test should be presented in the test manual.

> C3. The procedures for scoring the test should be presented in the test manual with a maximum of detail and clarity to reduce the likelihood of scoring error. (p. 18)

How much detail is required will vary with the complexity of the scoring procedure: Only minimal directions will be required for multiple-choice tests but detailed lists of correct responses and acceptable variations may be required on free-response tests (cf. the manuals for the Wechsler and Stanford-Binet Intelligence Scales). The directions are clear enough when they permit error-free scoring or, with free-response items, high agreement between scorers.

The manual should also suggest procedures that can be used to check for computational and clerical errors (C3.1) and give an estimate of the time required for scoring (C3.4). If there is more than one method of recording answers (e.g., in the test booklet or on a separate answer sheet), the manual should report data on the comparability of scores obtained by the various methods (C3.3).

On tests using free-response items (e.g., essays, short answers, oral responses), scoring becomes more complex and less objective; thus, more detailed scoring directions will be needed. In addition, the manual should present data on inter-scorer agreement (C3.2, C3.2.2). And, because use of such items usually requires special training of scorers, training procedures should be suggested (C3.2.1). On all types of items, it is also helpful for the manual to point out common scoring errors and suggest how they can be avoided.

The test user's primary responsibility is to ensure that the scoring procedures are carefully followed so that scoring errors are avoided.

> I3. A test user is responsible for accuracy in scoring, checking, coding, or recording test results. (p. 66)

If scoring is delegated to another person (e.g., a clerk), both the test user and the scorer share responsiblity for accurate scoring. In these cases, the test user should monitor the scoring to see that correct procedures are followed and verify the accuracy of scoring. If scoring requires some judgment, the test user should determine the degree of agreement between different scorers or the same scorer at different times (I3.2).

The previous comments have been directed primarily to situations where tests are hand scored. Often, however, tests are scored by electronic scoring machines. When such machines are used, the test user still has the responsibility for ensuring the accuracy of scoring.

> I3.1. When test scoring equipment is used, the test user should insist on evidence of its accuracy; when feasible, he should make spot checks against hand scoring or develop some other system of quality control. (p. 66)

Even though the current generation of scoring machines are more accurate than previous ones, they are not infallible. Thus, verification of accuracy by using other scoring methods is essential. If a commercial scoring system is used, the test user should insist that the scoring service provide data about the procedures used and scoring accuracy. Even if the error rate is low, spot checking should be done. Checking should include any level of scores (e.g., total scores, subtest scores, or scores on individual items) used when reporting performance or making decisions.

To summarize, the test publisher's responsibility is to provide detailed directions and procedures for scoring; the

test user's responsibility is to see that these procedures are conscientiously followed. To ensure this will occur, scorers will have to be trained to the point that they can apply the scoring procedures consistently and without error, understand why the procedures are as they are, and accept the need for accuracy in scoring. The reason, of course, is that use of erroneous scores will lead to improper decisions being made.

Other considerations. Before leaving our discussion, we will briefly consider two other topics related to test administration and scoring: the need for test security and the test taker's rights regarding his scores.

In many situations, results will be invalidated if test takers have knowledge about a test prior to taking it. A clear example is tests used for selection decisions, such as college and professional school admissions tests. But the problem of test security is much broader.

> Test security is a problem whenever a lapse in security can result in changing an individual's score without making a change in his true scores. ... Security may be compromised where examinees have had much prior experience with a popular test, have been taught specific test items, or have heard a lot about the test. (p. 67)

Stated somewhat differently:

> There are two principal reasons for controlling the use of psychological tests: (a) to prevent general familiarity with test content, which would invalidate the test and (b) to ensure that the test is used by a qualified examiner. (Anastasi, 1976, p. 30)

In other words, test security is necessary in order to ensure that the obtained scores accurately reflect the examinee's true performance and to ensure that the test is used properly.

The responsibility for maintaining test security rests on both the test publisher and the test user.

> I5. The test user shares with the test developer or distributor a responsibility for maintaining test security. (p. 67)
>
> I5.2. All reasonable precautions should be taken to safeguard test material. (pp. 67-68)

In practice, test publishers exercise their responsibility by requiring persons requesting secure tests to document their

qualifications (e.g., by selling tests only to persons employed in certain types of organizations or belonging to certain professional organizations) and/or by restricting testing to centers employing qualified personnel (e.g., the testing centers for college and professional school admissions tests). The test user can maintain security by such procedures as keeping tests in locked files and otherwise ensuring that unauthorized persons do not have access to the tests. But the problem of security goes beyond that of a potential test taker gaining access to a test; a well-meaning teacher who scans a test, then teaches students the content of specific test items can also destroy the value of the testing.

What if security is breached? If there is reason to believe that security has not been maintained, the test results should not be used as a basis for making decisions (I5.3). Instead, the test user should use other methods of assessment such as retesting with a different form of the test (if one is available) or using a different test or method of assessment (I5.1).

The second question is the test taker's right to know how he performed on the test, to know his score. The general principle here is that the test taker always has a right to know his score.

> J2.1. An individual tested (or his agent or guardian) has the right to know his score and the interpretations made. In some instances, even scores on individual items should be made known. (p. 69)

Note that this principle says that the individual has a right to know not only his score but the interpretations or inferences made from the scores. Thus the question becomes not should the examinee know his score, but how the score should be interpreted to him. This question will be discussed in more detail in Chapter 4.

Who else should have access to the scores? This is a difficult question to answer. If tests are used for decision-making purposes, then the person(s) who will make the decision obviously will need to know the scores. In some situations, such as vocational counseling, only the test user and test taker need to have access to the scores. Other situations are not so clear; for example, which teachers or school personnel should have access to students' scores on standardized achievement and ability tests? Whenever there is any question about who should have access to scores, a safe procedure is to have the test taker give his informed consent, in writing, before the scores are released.[4]

In all cases, scores should not be released to persons who are not qualified to properly use and interpret them.

> J2. Tests scores should ordinarily be reported only to people who are qualified to interpret them. If scores are reported, they should be accompanied by explanations sufficient for the receipient to interpret them correctly. (p. 68)

This means that reporting the score alone generally will not be sufficient; information regarding the circumstances under which the test was taken, conditions which may have influenced his performance, and information regarding the norm groups used as the basis for interpretation of scores may need to be provided.

One final question arises: What if a test taker has reason to question the accuracy of her scores? Certainly, the first step is to recheck the scoring for any errors. If the score is still in question, an opportunity for retesting, preferably with an alternative form of the test, should be afforded the test taker. If retesting is infeasible, alternative assessment methods might be used. In short, test takers should have the right to question their scores, ensure their accuracy, and be retested if questions remain unresolved.[5]

Summary

The points discussed in this chapter can be summarized briefly by four general principles. (1) When selecting an assessment procedure, consider the goals and objectives of the particular situation and all the relevant information regarding each potentially useful test. (2) Do not use a test or other assessment method you are not qualified to use. (3) In order to preserve the standardization of the test, follow the directions for administration and scoring exactly as they are presented in the test manual. (4) Consider the adoption of any test or assessment method as tentative: Constantly monitor and evaluate the testing procedures and the validity and usefulness of the assessment process.

Notes

1. Buros has also edited a number of other publications that present the same information as is in the Mental Measurements Yearbooks but whose content is restricted to a particular type of tests. Separate books are available for English tests, foreign language tests, intelligence tests, mathematics tests, personality tests, reading tests, science tests, social studies tests, and vocational tests. All are published by

Notes (Chapter 2, Cont'd)

The Gryphon Press.

2. Our comments, here and elsewhere, regarding comparisons between individuals apply particularly to norm-referenced tests, that is, to tests where scores are interpreted with reference to the test taker's relative position within some specified comparison group (the norm group). However, the need for standardized testing conditions applies equally as well to content-referenced tests where test takers' scores are interpreted with reference to a predetermined standard of content mastery. In the latter case, setting of the performance standard is based, at least partially, on knowledge of the conditions under which the test was taken. Thus, on all types of tests, when other than the standard testing conditions are used, interpretation of scores becomes problematical.

3. See Anastasi (1976, pp. 32-41) for an excellent discussion of these issues.

4. Government legislation (e.g., the Buckley amendment) also sets limits on the release of information about individuals, including test scores. See Chapter 5 for a more complete discussion of this issue.

5. The Educational Testing Service (ETS) has developed a series of exempliary procedures for dealing with bases of suspected cheating or incorrect scores (See Manning, 1978, p. 56, for a brief description of these procedures). Also, "truth in testing" legislation, such as recently passed in New York, specifies test takers' rights to knowledge about the test and their scores.

III. RELIABILITY AND VALIDITY

Besides becoming familiar with the development, standardization, and administration of a test, the user must also evaluate the characteristics of the test as a measuring instrument: its reliability, validity, and scores and norms. In this chapter we will consider reliability and validity; the discussion of scores and norms will be covered in the next chapter.

Reliability is concerned with the consistency of performance (i.e., test scores) over samples of items, test forms, and/or testing occasions. Unless performance is consistent we can have little faith in the predictions or inferences made from the test scores. This is clearly shown when reliability is thought of in terms of the amount of measurement error in obtained scores. If there is a large amount of measurement error - that is, scores are unstable or unreliable - scores will vary from situation to situation. These errors will, of course, make our predictions or inferences less accurate.

Validity is concerned with the accuracy of the inferences made from test scores.

> Questions of validity are questions of what may properly be inferred from a test score; validity refers to the appropriateness of inferences from test scores or other forms of assessment. The many types of validity questions can, for convenience, be reduced to two: (a) What can be inferred about what is being measured by the test? (b) What can be inferred about other behavior? (p. 25)

That is, validity evidence is designed to tell us what characteristic(s) the test measures and how accurately the test scores predict behavior in other situations.

As with other information about the test, evidence regarding reliability and validity should be presented in the test manual.

> B5. Evidence of validity and reliability, along with other relevant research data, should be presented in support of any claims being made. (p. 16)

It is not sufficient for the test developer and publisher to state the purposes and intended uses of the test (cf. B2, B3); they must also provide evidence to support these claims.

Test users have a similar responsibility - they, too, must present evidence to support their use of the test.

> E2. A test user is responsible for marshalling the evidence in support of his claims of validity and reliability. The use of test scores in decision rules should be supported by evidence. (p. 32)

When selecting a test, users should evaluate the validity and reliability evidence presented in the manual and in other sources. But this is not enough. As a test may behave differently in various situations, users should always collect evidence supporting the use of the test for their particular purposes in their particular situation. For example, there is extensive evidence that the validity of academic ability and college admissions tests (e.g., their ability to predict grades) varies from school to school. A test may be quite valid when used in one school yet have limited validity when used at another school. Thus, test users should always collect validity data in their specific situation.

Because of the technical nature of the concepts of reliability and validity, and their importance, the longest sections of the Standards are those concerned with these two concepts. We will discuss these concepts more briefly, looking at them from the test user's view. The reader who is going to conduct validity and reliability studies will need to study the Standards (and other sources) in more detail. Here we will focus on evaluating the validity and reliability information presented in the manual. Our goal is for you to be able to determine if the available information is adequate to justify your use of the test.

Reliability and Measurement Error

The concept of reliability refers to the degree to which the results of testing are consistent over forms or occasions; or, conversely, the degree to which the results are inconsistent due to certain sources of (measurement) error. Our goal is to obtain scores which are consistent from occasion to occasion and/or from one sample of items to another sample drawn from the same domain. Unless the test measures consistently (i.e., reliably) we cannot relate the test scores to other variables with any degree of confidence.

In many ways, reliability coefficients have limited

practical value for the test user. However, there are two places in the testing process where the user must be concerned with the reliability of the measuring instrument. One is during the test selection process. Here users should consider the evidence (for each prospective test) to see if the level of reliability is sufficiently high to justify use of the test.

> F1. The test manual or research report should present evidence of reliability, including estimates of the standard error of measurement, that permits the reader to judge whether scores are sufficiently dependable for the intended uses of the test. If any of the necessary evidence has not been collected, the absence of such information should be noted. (p. 50)

In practice this means that users should not use tests with low[1] or unreported reliability and, other things being equal, select the test with the highest reliability.

The other point in the testing process where reliability evidence becomes important is during the interpretation of test scores. Here our concern is somewhat different; in this case we want an index of how much individual scores will vary. This information is provided by the standard error of measurement. For the test user, this is often the most important use of the reliability data.

Types of reliability coefficients. There are several different types of reliability estimates. Most commonly, these are derived by correlating the scores of a sample of people on two sets of similar measurements; this correlation is called a reliability coefficient. As the different methods of estimating reliability take into account different sources of error, and because each method asks (and answers) a different question, they will not necessarily give similar estimates. Thus, test users should pay particular attention to those aspects of the reliability data that are most pertinent to their intended use of the test.

There are three sources of variability (error) that can influence test scores (Thorndike, 1951). The first set is concerned with the conditions of the test administration. Whenever testing conditions differ, an individual's scores are likely to vary, even if he takes the same test. For example, if we change the time limits, directions, or scoring procedures, or if the testing session is interrupted in some manner (e.g., by a commotion in the testing room), scores will be different than under other testing conditions. However, if the test administrator carefully follows the standardized instructions, these influences usually can be minimized.

A second source of error comes from the sampling of items comprising the test. Even when parallel forms[2] of a test are carefully constructed and standardized, it is often not possible to obtain completely equivalent forms. Furthermore, for any individual test taker, some items on one form may be easier or more difficult than comparable items on the other form. Thus, we would expect each individual's scores to vary somewhat, depending on the particular form of the test (i.e., sample of items) administered.

The third source of error, and the one that is hardest to control, includes those factors associated with the test taker. These may be relatively short-term fluctuations (e.g., momentary lapses in attention, reactions to specific items) or long-term changes in the individual (e.g., ones due to training, learning, or personality changes). Thus, even when the same test form is administered, if the testing sessions are separated in time there will be differential changes between test takers which will reduce the consistency of measurement and, thus, lower reliability.

These types of errors form the basis for a classification of types of reliability. In some situations we are interested in the stability of performance over time. For example, when using an interest inventory in counseling students regarding career choices, we assume that the clients' interests will not change drastically over time; that their interests at age, say, 35 will be similar to their interests at age 18. Evidence of test-retest reliability (also called the coefficient of stability) is germane here. In this situation our primary question is: How stable will scores be over time?

Retesting with the identical form of the test is particularly perilous when retesting after a short time period and/or with maximal performance tests. This is because the test taker may remember some of the items (and their answers), thereby producing a spurious correlation. Thus, the Standards recommends using a parallel form of the test for retesting, whenever possible. While this procedure eliminates the memory factor it introduces another possible source of error, noncomparability of items.

Regardless of whether the same test or a parallel form is used for retesting, the operational procedure is to correlate scores on the two test administrations. The test manual should report the results of such studies.

F6. The test manual should indicate to what extent test scores are stable, that is, how nearly the scores are likely to be if a parallel form of a test is administered

> after time has elapsed. The manual should also describe the effect of any such variation of the usefulness of the test. The time interval to be considered depends on the nature of the test and on what interpretation of the test scores is recommended. (p. 54)

This description should indicate whether the same or parallel form of the test was used (F6.1), the period of time between testings (F6.2), descriptive statistics for each testing (F6.2, F6.4), and a description of the factors that might have influenced the scores to change (F6.3, F6.3.1).

In other situations we are interested in the comparability of scores on two parallel forms of the tests; parallel forms being tests containing different samples of items selected from the same domain according to the same specifications. As use of parallel forms eliminates the memory factor, this procedure is useful in situations where repeated testing is necessary (e.g., pretesting and posttesting in a course, or when the results of the first testing are questioned). In these situations, the reliability estimate will be the correlation between scores on the parallel forms. This is called parallel forms reliability or, alternatively, the coefficient of equivalence. Here the primary question is: Would the test taker obtain the same relative score on two (or more) samples of items drawn from the same domain?

As with other types of reliability, the basic information should be presented in the manual.

> F4. If two or more forms of a test are published for use with the same examinees, information on means, variances, and characteristics of items in the forms should be reported in the test manual along with the coefficients of correlation among their scores. If necessary information is not provided, the test manual should warn the reader against assuming equivalence of scores. (p. 52)

Note that with equivalence estimates, the reliability coefficient is not sufficient by itself. The prospective user must be provided with data that allows him to determine if the forms of the test are, in fact, equivalent forms. This information will include a content analysis of the test items as well as descriptive statistics for all test forms. To illustrate why this information is needed, consider a case where (supposedly) equivalent forms were used as a pretest and posttest in a course. If the form used as the posttest was easier than the form used as a pretest, learning gains would be overestimated; conversely, if the posttest was harder, learning gains would be underestimated.

If parallel forms are administered separated by a time interval, we obtain a coefficient of stability and equivalence. In this situation both errors due to changes over time and ones due to the noncomparability of forms can effect the reliability estimate. Thus, this procedure generally gives the lowest estimate of reliability.

In some situations reliability must be estimated from a single administration of a test. In these cases a measure of the internal consistency of the test should be provided. Again, the relevant data should be in the manual.

> F5. Evidence of internal consistency should be reported for any unspeeded test. (p. 53)

Note that the standard refers to unspeeded tests. This is because internal consistency reliability estimates for speeded tests will be spuriously high; thus they should not be used nor reported (F5.2).

There are two situations where internal consistency estimates are used. One is when we have only one form of the test and it can be administered only one time. The typical classroom exam is a good example. Here the usual procedure is to divide the test into two equivalent parts, correlate scores on the parts, then correct the correlation for test length.[3] When using these procedures the main concern is the proper assignments of items to parts, either by matching on content or by a random-half procedure (F5.1, F5.4). Other methods of estimating reliability can also be used in this situation, e.g., analysis of variance.

The other situation where internal consistency estimates are used is when we are interested in the homogeneity of the test; that is, when the test items are viewed as a sample from a relatively homogeneous domain. Here we are asking whether the test items all measure the same characteristic, e.g., do all measure mathematical reasoning ability, verbal relations, or dominance. In this case, indices such as coefficient alpha and Kuder-Richardson 20 are used as measures of internal consistency. Whether or not such indices are appropriate will depend upon whether we are interested in determining if all test items were drawn from the same domain. For example, we would not expect high internal consistency on a survey achievement test as the test items will be drawn from various topical areas and may measure different skills (e.g., knowledge, application, analysis). In contrast, tests designed to measure a single psychological trait (e.g., numerical reasoning, spatial relations, or dominance) should exhibit high degrees of internal consistency.

Information regarding reliability studies. In addition to reporting reliability coefficients, the test manual should also describe the methods and procedures used to obtain these coefficients.

> F2. The procedures and samples used to determine reliability coefficients or standard errors of measurement should be described sufficiently to permit a user to judge the applicability of the data reported to the individuals or groups with which he is concerned. (p. 51)

As reliability will vary depending on the sample used (e.g., reliability coefficients will be higher in heterogenous than in homogenous samples, tests which are too hard or too easy for the people tested will have low reliability coefficients), the manual should describe the samples used in reliability studies (F2.1, F2.3). If coefficients are corrected for restriction of range, both corrected and uncorrected values should be reported (F2.2).

Generally, reliability estimates will be reported as either reliability coefficients and/or standard errors of measurement. If other measures are used they should be explained clearly.

> F3. Reports of reliability studies should ordinarily be expressed in the test manual in terms of variances of error components, standard errors of measurement, or product-moment reliability coefficients. Unfamiliar expressions of data should be clearly described, with references to their development. (p. 52)

While the Standards suggests use of variances of error components (in F3 and elsewhere), these indices are rarely used at the present time. However, the user should be aware that when they are used they may provide useful information on the sources of influence on test scores.

Many tests report more than one score; for example, subtest or composite scores may be reported as well as a total score. As subtest scores are based on fewer items than total scores, they will generally be less reliable. Thus, reliability estimates should be reported for any subtest scores that will be interpreted.

> F1.2. Standard errors of measurement and reliability coefficients should be provided for every score, subscore, or combination of scores (such as a sum, difference, or quotient) that is recommended by the test manual (either explicitly or implicitly) for other than merely tentative

or pilot use. (p. 50)

The manual should also report correlations between scores on parts or subsections of the test (F5.3) and between part and total scores (F5.3.1). In the latter case, the manual should call attention to the fact that these correlations will be inflated because the same items are included in both the part and total scores.

Content-referenced scores. The problems in determining the reliability of content-reference scores are somewhat different than those arising when using norm-referenced scores. The Standards makes no comments about the reliability of content-referenced scores. This is due to the fact that procedures were in the developmental stage at the time of publication and there was (and still is) little agreement as to how to determine the reliability of content-referenced scores. The reader is referred to Harris et al. (1974), Martuza (1977), or Millman (1974) for a discussion of the issues involved.

Errors in individual scores. The previous discussion related primarily to the test selection stage of the assessment process. But, as mentioned previously, the test user generally has a somewhat different concern: How consistent are individual test scores? This question is usually answered by the standard error of measurement, which is an index of the amount of measurement error in test scores.

The need to provide information on the standard error of measurement is mentioned frequently in the Standards (e.g., F1, F1.2, F2, F2.3, F2.3.1, F2.3.2, F6.4). A single overall estimate generally is not considered sufficient - estimates should be provided for various samples (F2.3, F2.3.1), for various score levels (F2.3.2), and for each score that will be interpreted (F1.2). These estimates may be derived from the reliability coefficients or based on empirical evidence.

For the user, the major implication is that scores should be considered only as "best estimates". That is, scores should be interpreted as bands or ranges, rather than as exact points on a continuum.

> J2.2.1. Scores should ordinarily be interpreted in light of their confidence intervals rather than as specific values alone. (p. 69)

This confidence interval will generally be some function of the standard error of measurement. Many test publishers facilitate this type of interpretation by reporting scores as

confidence ranges, either instead of or in addition to exact scores. For example, rather than reporting that Megan McGonigle scored at the 75th percentile rank on a reading test, her scores would be reported as falling in, say, the 70th-80th percentile rank band.

In addition to interpreting individual scores, there are at least three other situations where estimating the magnitude of (measurement) errors in scores is particularly important. One is when dealing with scores at, or near, the chance level of performance. These scores generally should not be interpreted (J5.5). Another is when considering differences between scores; say, for example, when interpreting differences between verbal and quantitative scores on an ability test, or between achievement and ability measures, or when using achievement test scores to measure growth from year to year. It is a well-known statistical fact that difference scores are less reliable than either of the individual scores. Thus, the manual should caution against possible overinterpretations of difference scores (F1.3), and indicate what size difference is needed for statistical significance (F1.4). Third, we can often tolerate a lower level of reliability if scores are to be used only for comparisons between groups rather than for interpreting individual scores. However, if the test is to be used for group comparisons, the manual should report data on the magnitude of the measurement errors for groups (F6.4).

Implications for the test user. We will conclude our discussion of reliability by reviewing the major implications of reliability data for the test user.

First, reliability is an important consideration when selecting a test. Other things being equal, the user should select a test with high reliability. Keep in mind that different types of reliability estimates do not provide the same information; they ask different questions and take into account different sources of error. Thus it is possible, for example, for test scores to be highly internally consistent yet unstable over time. Which particular estimate will be most important depends on how the test will be used.

Second, when interpreting test scores the major concern is generally with the amount of error in individual scores. Thus, the user must consider the standard error of measurement; the higher its value, the less confidence we can put in the score.

Finally, reliability should never be confused with validity. While an unreliable test can have only limited validity, just because a test is reliable does not guarantee that it

will be valid. The user should recognize this fact.

> J3. The test user should recognize that estimates of reliability do not indicate criterion-related validity. (p. 69)

This standard could be generalized to any type of validity; it is not restricted to criterion-related validity. Too often, test users assume that because a test is reliable and looks like it measures what it purports to measure, it will be valid and useful. Such is not the case. While reliability is a necessary condition for validity, it is not a sufficient condition.

Validity

If one characteristic of a test can be considered more important than any other it is validity. Questions of validity relate to what the test measures and/or predicts, what can be inferred from the test scores, and how useful the test results will be. These, of course, are the central questions when using any assessment procedure.

But what is meant by validity? The Standards gives a succinct answer to this question.

> Questions of validity are questions of what may properly be inferred from a test score; validity refers to the appropriateness of inferences from test scores or other forms of assessment. The many types of validity questions can, for convenience, be reduced to two: (a) What can be inferred about what is being measured by the test? (b) What can be inferred about other behavior?
>
> The first question inquires into the intrinsic nature of the measurement itself. The measuring instrument is an operational definition of a specified domain of skill or knowledge, or of a trait, of interest to the test developer or user. The essential problem in this context is to reach some conclusion as to how faithfully the scores represent that domain....
>
> The second question inquires into the usefulness of the measurement as an indicator of some other variable as a predictor of behavior. In this context, the essential problem is to reach some conclusion about how well the scores on the test are related to some other performance ... (p. 25)

Note the emphasis on validity being inferred. This means that

no single piece of evidence tells whether the test is valid:

> Validity coefficients may be presented in a manual, but validity for a particular aspect of test use is inferred from the collection of coefficients. It is, therefore, something that is judged as adequate, or marginal, or unsatisfactory. (p. 25)

In other words, users must consider the total body of validity evidence as it applies to their intended use of the instrument, then make a judgment of the adequacy of the validity evidence.

As with reliability, the Standards is primarily concerned with presentation of information about validity studies, rather than with how the test user can apply this information in a practical situation. Nevertheless, the user must be familiar with is information because:

> A test manual can provide evidence that will enable the user to evaluate the appropriateness of the item content, to determine whether the test is an acceptable measure of a specified construct, and to decide whether the test has provided useful predictive validities in situations similar to his own. An adequate research report can help the user decide whether to go ahead with the use of the test or to seek another predictor. (p. 31)

Therefore:

> E1. A manual or research report should present the evidence of validity for each type of inference for which use of the test is recommended. If validity for some suggested interpretation has not been investigated, that fact should be made clear. (p. 31)

Particularly important is the phrase "for each type of inference". As the Standards emphasizes: "No test is valid for all purposes or in all situations or for all groups of individuals" (p. 31). In other words, validity is situation-specific; all statements about validity should recognize this.

> E1.1. Statements about validity should refer to the validity of particular interpretations or of particular types of decision. (p. 31)

In brief, this means a test that is valid in one situation, or for one type of use, may not be valid in other situations or for other uses. Furthermore, these considerations apply to all scores which will be interpreted, be they total scores,

composite scores, subtest scores, difference scores, or profiles (E1.2, B5.5).

Users, in turn, have responsibility for providing evidence in support of their intended uses or interpretations of test scores.

> E2. A test user is responsible for marshalling the evidence in support of his claims of validity and reliability. The use of test scores in decisions rules should be supported by evidence. (p. 32)

This standard applies to two places in the testing process. When selecting a test the user has the responsibility to "read, understand, and evaluate the manual, the research, and the literature to show the appropriateness of the test for the intended use" (p. 32). This evaluation should also consider the rationale underlying the development of the test (A2, B2, B3), particularly as it relates to the user's goals and purposes (H1, H3). The other responsibility is to evaluate the test as it is being used. Whenever feasible, users should collect validity data on the test as used in their particular situation.

> E2.1. Test users are responsible for gathering data on the validity and reliability of their assessment techniques. (p. 33)

Realizing that this may be difficult or impossible in many situations (e.g., because of small sample sizes), the user should at least do informal evaluations and be alert to data suggesting lack of validity (E2.1). For example, if test scores are used to place students in various sections of a course and teachers report that many students appear to be improperly placed or there is a high drop-out rate, this might be a sign that the test scores are not providing the desired information.

Collection of new validity data is most important when a test is used in situations where it has not previously been validated (E2.2) and when the testing conditions have been altered significantly from the standard conditions (E2.3). An example of the former situation would be when a test used for selection (or placement) in one curriculum is used in another curriculum, especially one requiring somewhat different abilities and skills. An example of the latter would be if a test (e.g., an arithmetic computation test) that appeared too easy for a given group of students when given with the standard time limits was administered with shorter time limits.

Types of validity. The Standards recognizes three major types of validity: criterion-related validity, content validity, and construct validity. The distinction between types is based on the inferences drawn from the test scores. Before discussing each type in detail, we will give a brief overview of the three types.

Criterion-related validity is important when we want to infer (predict) an individual's standing on some other variable (the criterion) from their performance on the test. For example, we are interested in criterion-related validity when we predict performance in a foreign language course from scores on a verbal ability test or when we place students in various math courses on the basis of their scores on a mathematics ability test. The common element in all criterion-related validity studies is that test scores are used to predict performance in some other situation. Usually our primary concern is with the behavior to be predicted; scores on the test are of interest only because they indicate probable performance on the criterion measure.

There are two subtypes of criterion-related validity. One is predictive validity. Here we are interested in predicting behavior which occurs at some time in the future; that is, a time interval (and frequently some particular experience or training) intervenes between the test and collection of criterion data. An example is the use of college admissions tests (e.g., the ACT or SAT), taken in high school to predict college grades. In contrast, in concurrent validity studies the test and criterion data are collected at the same point in time. This information tells us whether one measure can be substituted for another (usually more expensive or less efficient) measure. For example, research has shown that scores on multiple-choice tests of correctness of expression are highly correlated with ratings derived from actual writing samples; thus, the more efficient measure (the multiple-choice test) can be substituted for the less efficient (writing sample) measure.

Content validity is important when we want to estimate how an individual would perform in the universe of situations (i.e., the domain) that the test items are designed to represent. Most frequently these are content or skill domains, but they also could be personality characteristics. Consider an achievement test covering high school biology. As we cannot administer all possible items to the students, we select a particular sample of items to comprise a test. Based on performance on this sample of items, we make inferences about their knowledge of high school biology (the domain). Needless to say, the crucial consideration in content validity is how

well the i s sampled by the test represent the broader domain.

Construct validity is appropriate when we want to more clearly define the trait or attribute (i.e., construct) measured by a particular test or assessment method. In psychology and education we use constructs such as verbal ability, reading readiness, anxiety, field independence, and hundreds of others. Operationally, these constructs are defined and measured by tests or other assessment methods. Our concern is whether, in fact, the measure actually reflects the underlying construct. The process of verification is called construct validation. This cannot be done by a single study; rather it requires accumulation of evidence from a variety of studies, evidence that clarifies the nature of the construct being measured.

Although we can distinguish three types of validity and, for convenience, can discuss them separately, they are not independent. Consider, for example, a test of mathematical ability. The concept of "mathematical ability" is a construct; a presumed characteristic on which people differ. Thus, we would need information to clarify the nature of this ability (trait); this would involve construct validation. When selecting items for the test, we would sample the various skills making up mathematical ability. The effectiveness of this sampling would be a question of content validity. Once the test is constructed, it is likely that we would use the scores to make inferences about performance in other situations, e.g., to predict grades in math courses or success in curricula or jobs requiring mathematical ability. Here criterion-related validity data would be needed. Thus, all three types of validity may be relevant for any test.

In educational settings we are generally most concerned with content validity. Educational programs are designed to teach certain content and skills and evaluation (either of the student, instructor, or method of instruction) involves determining how well students can perform the content and skills the class was designed to teach. As our major concern is with developing an instrument that adequately samples the learning domain, we are interested in the content validity of our measures. In other cases, particularly at points of transition between educational programs (e.g., between grades or between high school and college), criterion-related validity may be important. For example, whenever test results are used to select students for programs (e.g., admission to college) or place students in various courses or sections, criterion-related validity evidence is relevant. Construct validity is generally of lesser concern to educators; the exception being

tests which measure abilities or personality characteristics which influence the educational process (e.g., intelligence, verbal ability, test-taking anxiety, achievement motivation).

In the following sections we will consider each of the three major types of validity in more detail. We will emphasize those aspects of the Standards that are of particular relevance to the test user - that is, what validity evidence the user should pay particular attention to when selecting tests and the standards relating to evaluation of the test after it is put into use. We will not discuss, in any detail, the procedures for conducting validity studies; for this information the reader is directed elsewhere.[4]

Criterion-related validity. Data on criterion-related validity is relevant when scores on a test are used to make inferences about a person's performance on some other variable. In educational settings the two most common examples are predicting performance in an academic program (e.g., predicting grades or other outcomes in courses or curricula) and when assigning students to various sections of a course or to academic tracks. The former is a selection decision, the latter a placement decision. In both of these situations data on criterion-related validity can be used as a means of evaluating the adequacy of the decision-making process.

In either case the general validation procedure is the same. First, the test is administered to some sample of individuals. Then the treatment is applied (e.g., the students take the course). Third, some measure of the outcome of the program is obtained (the criterion measure). Fourth, the relationship between the test (the predictor) and the outcome (the criterion) is determined.[5] Finally, on the basis of this information, decisions are made as to whether to continue the assessment program as it now operates, revise it in some manner, or discontinue the program.

Let us consider the various components of this process. First, consider the criterion. In most, if not all cases, our primary concern is with the criterion performance. For example, in courses we are interested in students' mastery of the material taught. More specifically, we are interested in maximizing the learning of all students. Information on predictive validity helps by telling us how to maximize performance. In the selection paradigm, validity data tells us what types of students are most likely to profit from the (current) instruction and what student characteristics are related to performance. In the placement paradigm, the data tells us what sort of program maximizes learning for what type of student. In either case our major concern is with learning;

validity data are important because they give us clues on how to maximize learning.

If such information is to be of value we must have an appropriate criterion measure. Here a problem arises: In most situations there are various possible criterion measures. For example, in courses our goal is to maximize learning. But how do we measure learning? By grades? By scores on tests? If so, what type of tests; e.g., ones that measure retention of factual knowledge or ones that measure ability to apply that knowledge? What about performance in subsequent courses or on a job? And so on. Any study can only investigate one, or a small number, of the possible criteria. Furthermore, a test that predicts one criteria may not predict others; e.g., a measure may identify students who can successfully complete a program but fail to differentiate between people who complete the program with varying levels of success; medical school admissions tests may predict success in medical school courses but not clinical skills.

Not only must users recognize that there are various possible criteria, they must also consider the validity of the criterion measures. Any criterion measure must be relevant, that is, be an adequate representation of the behavior it is designed to measure. Thus the Standards recommends:

> E3. All measures of criteria should be described completely and accurately. The manual or research report should comment on the adequacy of a criterion. Whenever feasible, it should draw attention to significant aspects of performance that the criterion measure does not reflect and to irrelevant factors likely to affect it. (p. 33)
>
> E4. A criterion measure should itself be studied for evidence of validity and that evidence should be presented in the manual or report. (p. 34)

An elaboration of these standards with particular relevance for education is:

> E7.4.5. When validity for predicting grades in a course is reported, reasonably clear information should be provided regarding the types of performance required in the course, the nature of the instructional method, and the way in which performance is measured. If the test was administered after the course was started, this fact should be made clear. (p. 39)

Obviously, validation data collected on samples which are similar to the groups on which the test will be used are of more

value than data from dissimilar groups (E6.1.1). To enable the user to make this judgment, the manual should describe the composition of the validation samples, particularly in terms of variables that might affect validity, such as age, sex, socioeconomic status, and ethnic origin (E6.1). The manual should provide descriptive statistics - e.g., sample size, measures of central tendency and variability - for each sample (E6.2, E6.2.1, E6.2.2). Moreover, the manual should indicate the conditions under which the validation data were collected. For example:

> E6.1.2. If an ability test is to be used for educational or occupational selection, its validity should be established using subjects who are actual candidates and who are therefore ordinarily motivated to perform well. If the subjects used in a validity study are volunteers or were told that their test scores would not be used in making decisions about them, this fact should be made very clear. (p. 37)

The reason for this recommendation is, of course, that test takers' motivation will affect their performance and, hence, the validity estimates.

The other side of the sampling issue is the generalizability of the results of a validity study. The results of any single study describe the relation found between a test and a criterion in a particular situation under a particular set of conditions. It is unlikely that the user's situation will be identical to those of the validation studies. The question then arises: How widely can these results be generalized? Obviously there is no simple answer to this question. However:

> E5. The manual or research report should provide information on the appropriateness of or limits to the generalizability of validity information. (p. 35)

The best way to accomplish this is to present validity data from studies involving a variety of groups and criterion measures.

> E5.1. A test manual should report evidence of validity for each type of criterion about which a recommendation is made. If validity for some recommended interpretation has not been tested, this fact should be made clear. (pp. 35-36)

> E5.2. For any type of prediction, a test manual should report criterion-related validities for a variety of institutions or situations. Where validity studies have

> been confined to a limited range of situations, the manual should remind the reader of the risks involved in generalizing to other types of situations. (p. 36)

If a variety of data is presented, the test user can pay particular attention to data collected in those situations which are most similar to his own. In most educational uses, these will be schools having similar programs and student populations.

Given appropriate criteria and samples, the test user also has to be concerned with the manner in which the validation data were collected, analyzed, and reported.

> E7. The collection of data for a validity study should follow procedures consistent with the purposes of the study. (p. 38)

> E8. Any statistical analysis of criterion-related validity should be reported in the manual in a form that enables the reader to determine how much confidence is to be placed in judgments or predictions regarding the individual. (p. 40)

In terms of data collection procedures, the manual should indicate the nature of the sample (E6, E7.2, E8.3.2), possible contaminating factors (E7.3), the time elapsing between collection of the predictor and criterion data (E7.4), any training or experience that occurred during this time period (E7.4.4), and when the validity data were collected (E7.4.1). The Standards also cautions users against making long-term predictions, especially when they cover time periods longer than those used in the validity study (E7.4.2, E7.4.3).[6]

For the test user, the most useful information will be the indices of predictive accuracy and of the amount of error in the predictions (the standard error of estimate). Both should be reported in the manual. If validity coefficients are corrected for errors of measurement, both corrected and uncorrected values should be reported (E8.2, E8.2.1); while corrected correlations may have some theoretical usefulness, the uncorrected coefficients better indicate the practical usefulness of the test. When a test is suggested for use in making differential diagnoses, its validity for that purpose should be reported (E8.3.1). And, in situations where more than one predictor is used, the relationships between the variables and multiple correlations should be reported (E8.4, E8.5).

What sorts of predictive indices should be reported? The

Standards suggests four possibilities:

> E8.1. A report of criterion-related validity should give full information about the statistical analysis and should ordinarily include, in addition to such basic descriptive statistics as means and standard deviations, one or more of the following: (a) one or more correlation coefficients of a familiar type, (b) descriptions of the efficiency with which the test separates criterion groups, (c) expectancy tables, or (d) charts that graphically illustrate the relationship between test and criterion. (p. 40)

Validity coefficients are probably the most widely used indices. Besides their familiarity, they have certain other advantages: they describe the relationship throughout the entire range of predictor and criterion scores; they allow predictions of estimated criterion scores through use of regression equations; and they allow for computation of an index of prediction error, the standard error of estimate. Their main drawback is that they do not give a direct index of decision-making accuracy. Group separation indices are useful in many cases, especially for problems involving differential diagnoses, but have one weakness - with large sample sizes, small differences will often be statistically significant though useless for practical purposes, thus increasing the risk of overinterpreting the results. Either of the last two indices - expectancy tables or charts - will probably be of most use to persons interpreting test scores to individuals, as they graphically illustrate the relation of predictor and criterion scores in a readily understandable manner. It should be noted that there are other ways of reporting validity data that are not mentioned in the Standards but which may be useful to the decision-maker. A good example is an index of selection efficiency.[7]

Two final comments need to be made regarding criterion-related validity indices. First, the method of analysis should be appropriate to the nature of the data and the intended use of the test (E8.1.3). Different validation methods address different questions, thus are useful for different purposes. For example, while a validity coefficient indicates the strength of the relationship between predictor and criterion scores it does not tell how many correct decisions will be made using the test; a user interested in this information would need an index of selection efficiency. Second, validity coefficients, and other validity indices, should be cross-validated (E10). Data obtained on one sample always capitalize on the particular quirks of that set of data, thus may give an inflated estimate of the actual relationship. Cross-

validated indices, giving a more conservative estimate, guard against possible overinterpretations.

Another consideration, test fairness (or bias), is less straightforward. Yet, the user must always be alert to possible biasing factors in validity studies.

> E9. A test user should investigate the possibility of bias in tests or in test items. Whenever possible, there should be an investigation of possible differences in criterion-related validity for ethnic, sex, or other subsamples that can be identified when the test is given. The manual or research report should give the results for each subsample separately or report that no differences were found. (p. 43)

The concern with bias is not just a psychometric issue; civil rights legislation mandates such investigations whenever feasible. But the problem extends beyond concern with unfairness to the groups designated above, it encompasses all aspects of the testing selection; e.g., tests involving apparatus to be manipulated may be unfair to persons having certain physical limitations or characteristics, a test heavily weighted with reading skills may be unfair to a child who is bilingual or who has a reading disability. Thus, test users should always try to identify, and control, irrelevant variables which may introduce biases.

The problems of detecting bias are complicated by the varying definitions of fairness and lack of agreed-upon procedures for detecting bias.

> It is important to recognize that there are different definitions of fairness, and whether a given procedure is or is not fair may depend upon the definition accepted. Moreover, there are statistical and psychometric uncertainties about some of the sources of apparent differences in validity or regression. Unless a difference is observed on samples of substantial size, and unless there is a reasonably sound psychological or sociological theory upon which to explain an observed difference, the difference should be viewed with caution. (p. 44)

> Users should be aware, however, that a too-hasty acceptance of bias or of differential validity, if used in decision-making, may be as likely to produce unfair test use as is failure to consider the possibility. (p. 43)

We will have more to say about the test bias problem later (see Chapter 5). At this point, we caution you to be ever

alert for possible biasing factors but also not to look so hard for biasing factors that you see them even when none are present.

One final, though extremely important point: Test users should, whenever feasible, collect validity data based on the use of the assessment procedure in their particular situation.

> E11. To the extent feasible, a test user who intends to continue employing a test over a long period of time should develop procedures for gathering data for continued research. (p. 45).

The reasons for this recommendation are several and hopefully obvious. One, mentioned previously, is that the validity data reported in the manual will never be completely applicable to any specific new situation in which the test is used. Thus, local validity studies often provide more useful information (E5.2.2). Even when such data are collected, they may become obsolete due to changes in the population or programs. For example, a test used in a particular school may become less useful or valid when the school population changes (e.g., due to shifts in attendance boundaries or admissions procedures) or when the school curriculum undergoes a substantial change. In these cases, new validity studies will be needed. If at all possible, the test should not be used for making decisions while the validation study is being conducted (E7.1).[8] The results of such studies should be reported in revised editions of the test manual (E1.3) and in local research reports circulated to persons involved with the testing program.

Content validity. Most educational tests are designed to measure students' degree of mastery of certain prescribed content and/or skills. A test is a collection of items designed to sample these content and skill areas. On the basis of the students' performance on the test (sample of items), inferences are made about their mastery of the domain. Unless the items are a representative sample of the domain, these inferences will be in error. Content validation is designed to evaluate how well the sample of test items represents the domain of interest.

The basic logic of content validation is clearly set forth in the Standards:

> Evidence of content validity is required when the test user wishes to estimate how an individual performs in the universe of situations the test is intended to represent.

> To demonstrate the content validity of a set of test scores, one must show that the behaviors demonstrated in testing constitute a representative sample of behaviors to be exhibited in a desired performance domain. Definitions of the performance domain, the users' objectives, and the method of sampling are critical to claims of content validity. An investigation of content validity requires that the test developer or test user specify his objectives and carefully define the performance domain in light of those objectives. The definition should ordinarily specify the results of learning rather than the processes.... (p. 28)

Three aspects of this description are particularly important. First, the primary concern is with the outcomes of learning, not the processes by which the material was learned. Second, content validity is evaluated with regard to specified objectives. And, third, the validation process involves establishing that the sample of test items does accurately reflect the specified objectives.

The Standards includes only one major recommendation regarding content validity.

> E12. If test performance is to be interpreted as a representative sample of performance in a universe of situations, the test manual should give a clear definition of the universe represented and describe the procedures followed in the sampling from it. (p. 45)

Definition of the domain can be either by a listing of instructional objectives or a content-by-process outline (E12.2, E12.2.1). If statements are made relating the test content to courses of study or other sources of content (e.g., texts) these sources, and the date of preparation, should be listed (E12.3, E12.4). When experts have been used to judge the content validity of the items, their qualifications and the directions under which they made their judgments should be described (E12.1), as well as the degree of agreement between judges (E12.1.1). Possible biasing factors should be examined and reported (E12.1.2).

When evaluating evidence on content validity, the user should pay particular attention to the definition of the domain. This should be an operational definition, which describes the content and skills to be measured or the tasks to be completed. For example, the task might be described as factoring quadratic equations, writing a descriptive paragraph, identifying the causes of recessions, or sight-reading a musical piece. Note, that there is no need to specify the

psychological (cognitive) processes which enable the student to solve the problem, answer the question, or perform the task; the focus is on the tasks per se.

Given the definition of the domain, the user can evaluate how well the test items represent that domain. The manual should aid this process by providing information as to what skill is measured by each item.

> E12.2. In achievement tests of educational outcomes, the manual should report the classification system used for selecting items. (p. 46)
>
> E12.2.1. When an achievement test has been prepared according to a two-way content-by-process outline, that outline should be presented in the manual, with a list of the items identified with(in) each cell of the outline. (p. 46)

This latter recommendation could be expanded to apply to tests based on a list of instructional objectives; here the manual should list the instructional objectives and indicate which items measure attainment of each objective. In either case, the manual should provide the information necessary to connect individual items with the various components of the underlying content domain.

Given this information, users can judge whether the test has sufficient content validity for their purposes. This should be done even when evidence of content validity is provided in the manual because the test items have been selected to represent the domain as the test developer has defined it. That is, evaluations of content validity included in the manual will be with reference to the test developer's definition. What if the user's definition of the domain differs from the test developers? This problem arises whenever one uses standardized achievement tests. Standardized tests are designed to cover material commonly taught in American schools. Any given course or curriculum will have somewhat different emphases; a specific course will not cover all the areas measured by the test items, will include some topics not covered on the test, and will give different emphases to other topics. In this situation test users can describe the domain as they perceive it, then compare the test items to their definition of the domain. This procedure will give an indication of the content validity for their particular purposes. Needless to say, the closer the user's definition is to that of the test developer, the more valid, and hence the more useful, the test will be in her situation. This example also makes another important point about content validity - a test may lack content

validity either because the definition of the domain is inappropriate for the intended use of the test or because the items do not adequately reflect the domain as defined.

While we have discussed content validity in terms of paper-and-pencil achievement tests, the idea is also germane to other types of assessment, including personality measures. One example is the measurement of skilled performances (e.g., typing, driving, carpentry, or artistic skills), be they on the job or in courses. To evaluate the content validity of measures of skilled performance, one would first identify the component skills necessary to perform the task successfully, then determine whether the assessment instrument measured all, or a representative sample, of these skills.

Another example is ad hoc tests built to measure attainment of instructional objectives. A number of organizations have developed "banks" of test items designed to measure certain instructional objectives. The user specifies, generally from a standard list, those objectives which she wishes to measure. The test publisher then selects the required number of items from the item bank and constructs a test for that particular user. As different users will specify different combinations of objectives, each test will be an unique sample of items designed for a particular use. When evaluating the content validity of such tests, we must again recognize that the test will not have content validity unless the objectives specified reflect the major goals of the course and the items adequately measure mastery of the designated objectives.[9]

Finally, content validity should not be confused with face validity. Face validity means that the test items appear, on casual inspection, to be relevant for their intended purposes. Content validity, in contrast, is based on a systematic study of the relation of the test items to the description of the content/skill domain.

Construct validity. The goal of construct validation studies is to better describe (define) the trait or characteristic being measured by a test. In educational and psychological research and theory we use a variety of constructs (e.g., mathematical aptitude, verbal ability, creativity, anxiety, reading readiness) to describe traits or characteristics of individuals. When we measure a construct, by tests or other assessment procedures, we are essentially providing an operational definition of the construct. Our concern is both with the inferred construct and the measurement of the construct. Thus construct validity studies address two questions: (a) Is the test a good measure of the underlying construct? and (b) What is the nature of the construct?

Evidence of construct validation is not found in a single study; rather it derives from an accumulation of evidence. The procedure, in broad outline, is as follows. First the investigator formulates a tentative description of the construct. From this definition, he suggests a method of measuring the construct. He then makes certain hypotheses or predictions about how people who obtain different scores on the test (e.g., the measure of the construct) will respond to other measures and/or behave in other situations. These hypotheses are then tested through empirical studies. If the hypothesis is confirmed, there is some evidence of construct validity; if the hypothesis is disconfirmed, the investigator must revise his formulations. This revision may involve a redefinition of the construct or the theory surrounding it and/or a revision of the test to make it a better measure of the construct. Further studies are then conducted and:

> Through the process of successive verification, modification, or elimination of hypotheses, the investigator increases his understanding of the qualities measured by the test. Through the process of confirmation or disconfirmation, test revision, and new research on the revised instrument, he improves the usefulness of the test as a measure of a construct. (p. 30)

In short, the process of construct validation consists of continually formulating hypotheses about the test and construct, submitting them to empirical verification, and revising our ideas on the basis of the empirical results.[10] The goal is always to clarify our knowledge about the construct and the instrument devised to measure it.

When reporting information about construct validation studies, the test developer must state how he proposes to interpret scores on the measure of the construct.

> E13. If the author proposes to interpret scores on a test as measuring a theoretical variable (ability, trait, or attitude), his proposed interpretation should be fully stated. His theoretical construct should be distinguished from interpretations arising on the basis of other theories. (p. 46)

Note that the recommendation states that the manual should not only provide evidence in support of the proposed interpretation, it should also consider alternative interpretations of the results (see also E13.1, E13.2, E13.2.2, E13.2.6). Too often, measures of supposedly new constructs turn out, on careful inspection, to be nothing more than a previously defined variable masquerading under a new name.

Reports of construct validation studies should describe the operational procedures used in the study (E13.1) and the effects of other variables on the results (E13.2); for example, social desirability (E13.2.3), speed related factors (E13.2.4), and different test taking strategies (E13.2.5). Construct validation studies should also report the relationships between the test being studied and well-established measures (i.e., ones whose interpretation is relatively clear and accepted) of the same characteristic (E13.1.2, E13.2.1, E13.2.2). For example, scores on any proposed new measures of verbal and quantitative abilities should be correlated with scores on well-established measures of these variables. If the information obtained from the proposed test duplicates information obtained from an established measure, the new test will be of little value unless there are gains in the cost or efficiency of measurement. On the other hand, low correlations with established measures may (though do not necessarily) indicate a new variable has been identified and measured.

In summary, construct validation involves the accumulation of evidence about the nature of the trait under investigation, information that either clarifies the nature of the trait or extends the usefulness of the trait (or measure) to a wider variety of situations. In educational situations it is most useful with measures of characteristics of individuals (aptitudes, personality traits, attitudes) which may influence the educational process.

Implications for the test user. The implications for the test user of our discussion of validity should, by now, be clear. However, we will briefly review the most important points. First, the test user needs to consider validity evidence at two points in the testing process - when selecting a test for use and when evaluating its effectiveness once it is put into use. In both situations users must realize that no test will be valid for all purposes and for all groups. How valid the test will be depends both on the situation in which it is used (including the characteristics of the population tested) and the types of inferences made from the test scores. A test that is valid for one situation or use may not be valid in another (even presumably similar) situation or use. Thus, test users must always evaluate the validity of a test for their purposes in their situation. If at all feasible, this requires collecting validity data in the user's particular situation.

Second, the user should recognize that there are several types of validity, each of which is designed to answer a different question. Which particular type, or types, will be

most relevant depends on the situation and how the test will be used (i.e., what inferences will be drawn from the test scores). In both the test selection and the test use aspects of the assessment process, users should pay particular attention to those types of validity evidence which are most relevant to their use of the test.

Third, validation is not a one-shot, yes-or-no process; it is a continual process of obtaining more information about the nature of the test and what behaviors can be predicted by the test scores.

Summary

In this chapter we have considered two properties of tests, reliability and validity. Both are concerned with the influence of certain variables on test scores. Reliability is concerned with identifying and estimating the magnitude of those effects which cause scores to vary and be inconsistent. Validity, in contrast, is concerned with the effects of those variables that cause a test to be a less useful or clear measure of the characteristic(s) it is designed to measure.

For the test user, reliability information often is of limited value. In most situations, the user's concern is only to ascertain that the measure has sufficient reliability for its intended use. However, if scores will be interpreted to individuals, the user must also be concerned with the magnitude of the standard error of measurement. Users should avoid the common error of assuming that a test which is reliable will also necessarily be valid. While reliability is a prerequisite for validity, a highly reliable test may not measure the characteristic intended nor predict any relevant outcomes.

Validity is the single most important property of a test. Unless we have validity evidence, we do not know what the test measures or how well it can predict other behaviors. As there are various types of validity, the user must look to the evidence most relevant for his particular use and situation. The responsibility is on the user to justify his use of the test. For example:

> I4. If specific cutting scores are to be used as a basis for decisions, a test user should have a rationale, justification, or explanation of the cutting scores adopted. (p. 66)

More often than not this means collecting validity data in the situation in which the test will actually be used.

Notes

1. Because of the various types and uses of tests, and the different types of reliability coefficients, it is impossible to specify a particular value as indicating that a test has attained a satisfactory level of reliability. However, reliability coefficients on many standardized maximal performance tests are .85-.90 or higher and frequently as high as .95 for composite scores. Tests with lower reliabilities should be considered with a degree of skepticism, unless we plan to interpret only group means (and not individual scores). In these cases, a lower level of reliability may be tolerated.

2. Parallel forms of a test are two (or more) samples of items selected from the same domain according to the same specifications. The forms can be considered equivalent if, in addition to having similar content, they also have comparable means, variability, and distributions.

3. This split is usually done post facto, rather than administering the test in two sections. A common practice is to use the number correct on the odd-numbered items as one score and the number correct on the even-numbered items as the other score. A correction for test length can be made using the Spearman-Brown formula.

4. A good source is Principles for the validation and use of personnel selection procedures (APA Division 14, 1975). For more general discussions of validity see, e.g., Anastasi (1976, chaps. 6 & 7), Brown (1976, chaps. 4, 7 & 8), Cronbach (1970, chaps. 5 & 13), Cronbach (1971), Ebel (1972, chaps. 14 & 16), or Kerlinger (1973, chap. 27).

5. More than one predictor and/or criterion measure can be used in a study. For simplicity, we will discuss only the case using one predictor and one criterion measure.

6. For example, most studies relating scores on college admissions tests to college grades use grades obtained during the first term or first year as a criterion measure. To generalize predictions beyond this period is risky as evidence indicates that the relationship decreases over time (see, e.g., Humphreys, 1968).

7. Indices of selection efficiency (or decision-making accuracy) indicate how many correct decisions will be made if a particular selection strategy or cutting score is used. For a discussion of these indices see Brown (1976, pp. 109-117) or Meehl & Rosen (1955).

8. This is both to ensure against use of unvalidated measures and to avoid contamination of the data in the validity study. A related example illustrates how this can be accomplished. Publishers of the SAT include trial test items among scored items on regular administrations of the test. (Test takers do not know which are trial items and which ones are used to derive their test scores.) Data from these trial administrations are then used to select items for inclusion on further revisions of the test.

9. The Standards does not discuss the validity of mastery (content-referenced) tests. Here two questions are of interest: How valid is the test? What score level indicates mastery? Readers interested in these questions are referred to Hambleton et al. (1978), Harris et al. (1974), Martuza (1977), or Millman (1974).

10. For example, construct validation of a paper-and-pencil measure of test anxiety might involve correlating scores on the test with physiological indices of anxiety, finding out if students with high scores on the test score lower on ability tests when the testing situation involves some pressure (e.g., scores will be used to make decisions about them) than under neutral conditions (e.g., the test is administered as part of a research project), and seeing if training in test-taking skills and relaxation techniques lowers scores on the test anxiety measure.

IV. INTERPRETING TEST SCORES

Perhaps the most important aspect of the testing process is the interpretation of the test scores, for here the results of our assessment are put into use. Even if an instrument has been carefully constructed and standardized, and even if it is reliable and valid, unless the scores are properly interpreted our efforts will have been in vain. For at this point our assessments touch the lives of the people being assessed.

> ... decisions may profoundly influence the lives of those tested, such as decisions for employment or for attendance at college, or decisions to assign a person to one treatment or opportunity rather than to another (e.g., tracking in a school system), or decisions to continue or terminate a program or to regulate its funds. The cost of error, in money and in human suffering, may be great. (p. 57)

While these comments relate to decision-making uses of tests, similar comments could be made about descriptive uses. For example, a student may consider a counselor's description of his abilities when making decisions about post-high school plans such as which colleges to apply to or what sort of employment to seek. If the counselor's description leaves an erroneous impression with the student, he may make a wrong decision. Thus, test interpretations must be done carefully.

In this chapter we will first discuss where to find the information necessary for proper interpretation of scores, next consider the various types of scores, then discuss the process of interpreting scores. When reading this chapter, remember that all standards will not apply in every situation; you should always consider the particular situation and the purposes of the testing. To cite but one example:

> H5. Test scores used for selection or other administrative decisions about an individual may not be useful for individual or program evaluation and vice versa. (p. 63)

That is, how scores are interpreted will vary depending on the purposes of the testing. This may raise problems, as the

comment following H5 indicates:

> The purposes of institutional testing and of evaluative testing are not always compatible. Whereas the typical evaluative use is intended to help the individual (or program), institutional decisions frequently have the apparent effect of hurting, even if the decision may have unseen long-term benefits to the individual (such as avoidance of an unnecessary failure experience). These seemingly contradictory functions may prevent effective interpretation in either instance. (pp. 63-64).

This comment also makes another important point: We must always consider not only the correctness of the interpretation but also the effect that the interpretation will have on the test taker.

Background Considerations

As with other aspects of assessment, the responsibility for proper interpretation of test results is shared by the test publisher and the test user. The test publisher's responsibility is to provide the user with the information necessary to interpret the test properly. The test user's responsibilities are to ensure that he is properly qualified to interpret the test and to communicate the results in understandable terms to the test taker and/or other persons using the test results.

Information about scores. The information needed to interpret scores correctly should be found in the test manual.

> B1. The test, the manual, the record forms, and other accompanying material should help users make correct interpretations of the test results and should warn against common misuses. (p. 13)

This obligation extends beyond just providing technically correct information on such matters as how scores are derived and characteristics of the norm groups (see below). It includes such requirements as selecting names for the test, or its parts, which minimize the risk of misinterpretation (B1.1, B1.1.1); drawing the user's attention to data that must be taken into account when interpreting test scores (B1.2); calling attention to factors known to influence the test scores (B1.3); and drawing attention to, and warning against, interpretation errors that are known to occur frequently (B1.4).

Some examples may clarify these points. Using terms such as "intelligence", "culture-free", or "creativity" in test

titles, without clearly defining these terms or giving supporting evidence, may mislead the test user. Achievement tests requiring a high level of reading ability may measure reading skills more than knowledge of the subject area. Sometimes performance is influenced by the students' sex, physical limitations, or socio-economic background. And, as an example of an interpretation error, many students, and some test users, think scores on interest inventories are indications of ability or potention success in an area.

As the interpretation of scores may vary with the purpose of the testing, and the psychological and/or educational rationale underlying the test, the test manual should include statements about these points (B2, B3). The test user should take these factors into account when interpreting scores. Furthermore, if there is background literature relevant to the test and its interpretation, the manual should direct user's attention to these references (B4.3).

Finally, information should be provided for all scores which will be interpreted (B5.5). Data on total or composite scores are not sufficient. If subtest or scale scores are recommended for use (other than just for research purposes), enough information should be provided in the manual so that these scores can be interpreted accurately.

Qualifications of the test interpreter. All tests cannot be interpreted by all persons. As we mentioned previously, tests can be categorized in terms of the qualifications of their users. What these qualifications are should be indicated in the test manual.

> B4. The test manual should identify any special qualifications required to administer the test and interpret it properly. (p. 15)

These qualifications include knowledge of the principles of educational and psychological measurement (G1, G3) as well as knowledge of the literature relevant to the particular test and its uses (G2).

But tests are not administered and interpreted in a vacuum: The test interpreter must understand and consider factors relevant to the particular situation in which the test is used and the characteristics of the person who will receive the test interpretation. This requires a broad understanding of psychology and education plus a large does of common sense. For example, when interpreting scores from standardized achievement tests used in a particular school, the test user must consider the nature of the school and

community, the goals and objectives of the educational system, and the characteristics and goals of the students who were tested. To illustrate, if the purposes of the testing were not clearly explained to the test takers, or for some other reason students were not motivated to exhibit their best performance, the results may not reflect the students' true abilities. Unless such factors are considered, the interpretation may be technically correct (in this case, indicate how the students performed under these testing conditions) yet be of little value, or worse yet, misleading.

A caveat. The interpretation of test scores is usually the final step in the testing process. Thus, it builds on all the previous steps in the test development process. Of particular importance is the fact that the potential usefulness of scores depends on the reliability and validity of the test (which, in turn, reflects the adequacy of the test development and standardization processes). Too often, test users assume that because normative or other interpretive data are provided, scores will be meaningful and of value. Such is not the case. Unless there is evidence that the test is reliable and valid, the scores may be in error. The moral is clear: Before interpreting scores the test user must be assured that the test is a valid measuring instrument for the intended purpose.

Types of Test Score Interpretations

Just as there are various types of tests (and validity and reliability indices), there are various types of test scores. We will be concerned with three major types of test scores: norm-referenced scores, content-referenced scores, and criterion-referenced scores. Each class, in turn, contains various specific types of scores.

As norm-referenced scores have been widely used, they are familiar to most people. With norm-referenced scores, interpretations are made with reference to the performance of other people. That is, individual scores are interpreted in terms of their relative ranking within some specified comparison group (the norm group). Percentile ranks are a clear example of norm-referenced scores: A percentile rank of 72 indicates that the test taker obtained a higher score than 72% of the people in the normative sample.

In contrast, content-referenced scores are interpreted in terms of the level of performance on the achievement continuum measured by the test. Content-referenced scores differ from norm-referenced scores in that their frame of reference is the content domain of the test, not the performance of

other test takers. When we say that a student can correctly spell 75% of a standard list of spelling words we are making a content-referenced score interpretation.

The third type of score, criterion-referenced, is probably least common. In a criterion-referenced score interpretation, scores are interpreted in terms of performance on some external variable. For example, when we interpret a student's score on a college admissions test as indicating that she will probably obtain a B average in college, we are making a criterion-referenced interpretation as we have expressed performance in terms of an external variable (college grade average). Expectancy tables are another example of criterion-referenced test score interpretation.

There is some confusion over use of the term criterion-referenced scores. This term has been used by many people, particularly in educational testing (e.g., Glaser, 1963; Popham & Husek, 1969), to refer to criterion-referenced and/or content-referenced scores. Because the term "criterion" has a well-established meaning in terms of validation, we will use the terms as they have been defined above (i.e., as they are used in the Standards).

Although we have referred to norm-referenced, content-referenced, and criterion-referenced scores, it would be more correct to refer to these classifications as types of test score interpretations. This distinction, though subtle, is important because it indicates that the difference is in how scores are interpreted, not in the scores per se. This difference can be clarified by an example. Consider the score that a student makes on the Psychology achievement test of the Graduate Record Exam. This same score could be interpreted in terms of how well the student performed in comparison to other test takers (a norm-referenced interpretation), as an indication of her degree of knowledge of the material taught in undergraduate psychology courses(a content-referenced interpretation), and/or, given the necessary validity data, as a prediction of her probable success in graduate work in Psychology (a criterion-referenced interpretation). This example points out that the distinction between types is primarily in terms of how the scores are used and interpreted, not in the scores themselves. It also points out that the same score can be interpeted in all three ways.

Norm-Referenced Score Interpretation

Traditionally, most scores used on psychological and educational tests have been norm-referenced. That is, the test taker's performance is compared to the performance of other

people comprising some well-defined norm groups.[1] When tests are used for decision-making purposes, this group usually consists of potential competitors. For example, on college admissions tests the appropriate norm group would be other students applying to the same college or type of colleges. When tests are used for descriptive purposes, the norm group generally consists of people sharing some salient characteristic(s) with the test taker. Thus, the scores of a fourth grade student on a standardized reading test would be compared to other fourth graders in the same school system, community, state, and/or to a representative national sample of fourth grade students.

In order to interpret norm-referenced scores, two conditions must be present. The first is data be available on the performance of (a representative sample of) the comparison group. As noted above, these will be people who share certain characteristics with the test taker. Depending on the nature of the test and the purpose of the testing, these characteristics may include any or all of the following: age, sex, race, educational (grade) level, specific educational preparation, type of school, socio-economic background, primary language, community, or geographic region. In some instances, the group may be defined primarily by how the test is used; e.g., the common element defining the norm groups on college admissions tests is that all test takers are applying for admission to colleges or universities. In all cases, however, the basic requirement is to obtain a distribution of scores for the people comprising the norm group.

The second essential condition is the method of reporting performance, the score itself. The score, in some way, must indicate the test taker's relative position within the norm group. Common examples of norm-referenced scores are percentile ranks, standard scores (in all their forms), grade-equivalent scores, and intelligence quotients (IQs).

Norm groups. As norm-referenced score interpretations are based on a comparison of an individual test taker's score to some specified comparison group, knowledge of the nature and composition of this norm group is essential for accurate score interpretation. The overriding consideration is that scores be interpreted with reference to an appropriate comparison group or groups:

> J5. In norm-referenced interpretations, a test user should interpret an obtained score with reference to sets of norms appropriate for the individual tested and for the intended use. (p. 70)

As the Standards points out, the reverse is also true: We should never interpret scores with reference to a norm group that is inappropriate for the individual tested or the intended use of the test.

To decide whether a particular set of norms is appropriate (or inappropriate) requires comparing the characteristics of the test taker to those of the members of the norm group.

> J5.3. A test user should examine differences between characteristics of a person tested and those of the population on whom the test was developed or norms developed. His responsibility includes deciding whether the differences are so great that the test should not be used for that person. (p. 71)

But what characteristics? Here we are concerned with those characteristics described above, e.g., age, sex, educational level. Which ones will be important, and which will be of little concern, will vary from situation to situation. For example, when interpreting scores of elementary school children, grade level, previous educational background, and type of school will obviously be important factors. That is, we would (generally) want to compare students to other students in the same grade, in the same type of school, who had similar backgrounds in the subject matter being tested (e.g., compare students taking "new math" with others having been taught new math rather than those who had "traditional" math courses).

In most situations, which of the defining variables are of concern will be relatively clear. However, two variables - race and sex - are more controversial. Some people think that the treatment of the different sexes and races in our culture is such that these factors should be taken into account when interpreting test scores by using separate norm groups for males and females and/or members of different racial groups. There are others who consider these differences to be less crucial; for example, some people think that the crucial variable is socio-economic background or skill and experience in the use of Standard English rather than race. Unfortunately, the available data often do not provide a clear-cut answer. Thus, many tests have separate norms for males and females and/or for different racial groups. In these situations the test user can do either of two things. One possibility is to look at the differences in performance between the groups and determine if the differences are large enough to justify use of separate norm groups (or whether scores can be interpreted using a composite norm group). The other possibility is to interpret scores with reference to both norm groups. For example, scores of a black student might be compared to both black and

white norm groups. When doing this users should be aware that they will obtain different information from each comparison. By comparing performance to other blacks, we will get an indication of how the student compares to other students with (presumably) comparable backgrounds; by comparing scores to the white norm group we will get an indication of how the person's scores compare with people in the dominant (i.e., white) cultural group.

Like other standardization data, normative data should be included in the test manual.

> D1. Norms should be published in the test manual at the time of release of the test for operational use. (p. 20)

The most crucial information is a description of the characteristics of the population(s) that form the basis of the norm group(s) and how people were sampled to form the norm group.

> D2. Norms presented in the test manual should refer to defined and clearly described populations. These populations should be the groups with whom users of the test will ordinarily wish to compare the persons tested. (p. 20)
>
> D2.2. The description of the norms group in the test manual should be complete enough so that the user can judge its appropriateness for his use. The description should include number of cases, classified by one or more of such relevant variables as ethnic mix, socioeconomic level, age, sex, locale, and educational status. If cluster sampling is employed, the description of the norms group should state the number of separate groups tested. (p. 21)

These descriptions should be "clearly and prominently described in the manual" (D2.2.1, p. 21).

Because the norm groups will generally be based on a sample of the target population, the sampling process should also be described in the manual. This description should include discussion of any possible biases in the sampling procedures (D2.1.2), the number of sampling units and the number of individuals within each unit (D2.1.3, D2.3), and the conditions under which the normative data were obtained (D2.5). If the sample is small, or otherwise undependable, test users should be cautioned about the magnitude of possible errors in the interpretation of scores (D2.3).

If the test results are to be used to evaluate groups, rather than or in addition to individuals, normative data based on group statistics should be included in the manual.

> D7. Where it is expected that a test will be used to assess groups rather than individuals (i.e., for schools or programs), normative data based on group summary statistics should be provided. (p. 24)

An example would be when one wanted to compare the achievement or ability level of students in various schools. Here school averages would be the data of interest. Some test publishers provide normative data, based on group averages, for students enrolling at various types of schools or colleges.

As with other types of information, our concern is not just with the technical accuracy of the material but also with the manner in which the information is presented. Two aspects are of particular concern. One is the generality of the normative data:

> D2.1. Care should be taken to avoid misleading impression about the generality of normative data. (p. 20)

This caution is necessary because:

> Truly representative national norms, for example, are rarely if ever obtained; normative data collected from people or schools with specific characteristics, however, are frequently used as if they were taken from a representative national group. ... It is an error of interpretation to assume that the norms of the volunteer group of schools apply to schools in general; the incidence of such errors may be reduced by manuals that clearly define the characteristics of the normative populations. (pp. 20-21)

That is, as normative data often are not completely representative due to the voluntary nature of schools' participation, the user should be cautious in generalizing beyond the particular sample on which the norms were based.

The other point of concern is with subtest scores. Normative data should be reported separately for any scores that will be interpreted. Because subtest scores will be based on a smaller number of items than total scores, the scores probably will be less reliable and/or valid. Thus:

> D2.4. Norms on subtests or groups of test items should be reported in the test manual only if the validity and

> reliability of such subtests or groups of items are also indicated. (p. 22)

Too frequently the implication is left with the test user that subtest scores are equally as reliable and valid as total scores and can, therefore, be interpreted with the same degree of confidence. As pointed out in D2.4., this is not necessarily the case.

In summary, comprehensive information on the development and composition of the norm groups should be reported in the test manual. The test user's responsibility is to consider this information when making score interpretations, being careful not to generalize beyond the limitations of the available normative data.

Local norms. In many situations, a test user is more interested in comparisons with some circumscribed local population than with a broader group, such as a national sample. For example, a teacher may be more interested in comparing the performance of her students on a reading or arithmetic test to other students in the same school, or school system, than to students across the country. In these cases, local norms will be appropriate.

> D4. Local norms are more important for many uses of tests than are published norms. A test manual should suggest using local norms in such situations. (p. 22)

As the test publisher cannot be expected to develop local norms for each user, test users may have to develop their own normative data.[2] However, some test publishers will provide local norms as a part of their scoring services.

Local norms are particularly important when the test is being used for decision-making purposes.

> J5.4. Local normative data or expectancy tables should ordinarily be developed, if possible, when administrative decisions are based on test scores. (p. 71)

The reasons for this recommendation are obvious: the decisions will affect only people in the local population. To continue the example above, if the reading (or arithmetic) scores are used to place students in tracks or courses within a particular school, such groupings should be made on the basis of comparisons with local students; how students in other schools perform is irrelevant. Similar considerations would hold for other decisions such as identifying students needing remedial instruction or selecting students for

admission to a particular college or university.

On the other hand, even when a test is used primarily for local purposes, national normative data may provide useful information. One situation is when, because of small local samples or other reasons, use of local norms is not feasible.

> D1.2. Even though a test is expected to be used primarily with local norms, the test manual should nevertheless provide normative data to aid the interpreter who lacks local norms. (p. 20)

Such instances frequently occur when the local population is small and when testing students from minority groups or students who have specific handicaps or limitations.

In other situations the test user may have local normative data available but also want to make comparisons to a broader (e.g., a national) sample. For example, a teacher may want to compare students' performance on a reading test both to their classmates (local norms) and to a national sample (general population norms). Such comparisons will be most useful when the characteristics of the local sample differ, on some crucial dimensions, from those of the general population. For example, students from a highly selective or upper socioeconomic level high school will probably obtain lower percentile ranks on an achievement or ability test when compared to a local norm group than when compared to a national norm group. To interpret the students' scores only with reference to other students within the local system could give a misleading picture of their abilities.

There is also no reason why the same score cannot be interpreted with reference to several different norm groups. In fact, in some situations this would be desirable. The example in the previous paragraph is one such instance. Another example is with college admissions tests. Here the student would want to compare her scores with those of students at each of the colleges she was considering. Furthermore, if data were available, she might even want to compare her scores to students in various curricula within each college.

What if no appropriate normative data are available or can be developed? This is most likely to occur when testing students from a particular subsample of the population such as students from a minority group or students having a specific limitation or disability (e.g., blind students). Here the Standards recommends "doing the best one can":

> J5.3.1. If no standardized approach to the desired

> measurement or assessment is available that is appropriate for a given individual (e.g., a child of Spanish-speaking migrant workers), the test user should employ a broad-based approach to assessment using as many methods as are available to him. (p. 71)

This might include using a variety of assessment instruments, administering the test as a method of observing how the test taker attacks the problems, or even, in some cases, looking at the available norms with the realization that they are far from perfect. The point is that we should not, and cannot, abandon our efforts to solve educational or psychological problems just because appropriate normative data are not available.

Types of norm-referenced scores. Obtaining normative data is one essential element in interpreting performance on a norm-referenced test. The second step generally is to transform the raw scores to another scale which more directly shows the person's relative position within the norm group. This is not the place to describe the construction and interpretation of the various types of norm-referenced scores.[3] Rather we will discuss some general considerations in the use of norm-referenced scores.

The goal when interpreting norm-referenced scores (or any other type of score, for that matter) is to present the information obtained from the test in a manner that is technically accurate, understandable by the test taker or score user, and avoids possible misunderstandings. To accomplish these ends, the Standards recommends use of only certain types of scores.

> D3. In reporting norms, test manuals should use percentiles for one or more appropriate reference groups or standard scores for which the basis is clearly set forth; any exceptional type of score or unit should be explained and justified. Measures of central tendency and variability always should be reported. (p. 22)

> J5.1. It is usually better to interpret scores with reference to a specified norms group in terms of percentile ranks or standard scores than to use terms like IQ or grade equivalents that may falsely imply a fully representative or national norms group. (p. 70)

> J5.2. Test users should avoid the use of terms such as IQ, IQ equivalent, or grade equivalent where other terms provide more meaningful interpretations of a score. (p. 70)

The message should be clear: The preferred types of norm-referenced scores are percentile ranks and standard scores.

Percentile ranks generally cause little trouble, either to the test user or persons to whom the scores are being interpreted. The only caution is to recognize that a percentile rank score refers to the percent of people rather than percent of items; i.e., a percentile rank of 56 indicates that the test taker scored higher than 56% of the people in the norm group, not that 56% of the items were answered correctly. Standard scores, being more complex statistically and probably less familiar, are likely to cause more trouble, both to the test user and to the recipient of the interpretation. Thus:

> D5. Derived scales used for reporting scores should be carefully described in the test manual to increase the likelihood of accurate interpretation of scores by both the test interpreter and the examineer. (p. 22)

The manual should clearly describe the particular type of transformation used (D5.1, D5.2), whether standard scores are linear or normalized transformations (D5.2.1), and how much measurement error is involved in the scores (D5.2.2, D5.4). Such information is important because there are various types of standard scores, each of which differs in its computation. For example, both the ACT and SAT report performance on a standard score scale but the former uses a scale with a mean of 20 and a standard deviation of 5 points while the latter uses a scale with a mean of 500 and a standard deviation of 100 points. And, while both the Stanford-Binet and WISC-R IQs have a mean of 100, their standard deviations are 16 and 15 points, respectively.

Because standard scores have certain statistical advantages, they are often preferred to percentile ranks by test developers and users. However, because they are harder to understand, many people transform standard scores to percentile ranks when interpreting scores. For example, if the scores are normally distributed, a standard score of 60 (on a scale having a mean of 50 and standard deviation of 10) could be interpreted as a percentile rank of 84. The advantages of both approaches can be retained if score profiles are plotted on normalized percentile rank charts (D5.3).

The reader should also note that the Standards recommends that certain types of scores - IQs, IQ equivalents, mental ages, and grade equivalents - should be avoided if at all possible (D5.2.3, J5.1, J5.2). The reasons are several. One is the technical inadequacies of such scores. For example, derivations of grade-equivalent scores often require

interpolations to obtain scores at the extremes of the distribution; thus, extreme scores may contain a large amount of error and lead to improper interpretations. An equally important consideration is that such scores are frequently misinterpreted. For example, many people have preconceived ideas about what is a "good" IQ score and/or interpret the IQ as a measure of an individual's worth. Or, a parent, finding that Mike's reading grade equivalent score is 4.9 when he was tested in the second month of fifth grade may (erroneously) conclude that his reading abilities are deficient when, in fact, he is in the average range for children in his grade. Thus, use of these types of scores should be avoided. If the test user feels he must use these scores, it is a good policy to also interpret the scores in terms of percentile ranks (and/or standard scores) with reference to a well-defined norm group (see comment under J5.2, pp. 70-71).

Summary. Norm-referenced test score interpretation involves describing test performance in terms of a relative ranking within a specified population. To interpret such scores requires having information on the performance of this population (the norm group) and a score that describes the test taker's relative ranking within this population. The information on the development of the norm groups and scores should be found in the manual. If no appropriate normative data are available, the user may have to construct local norms. The user must become familiar with the normative data, then interpret performance in a manner that is both technically accurate and understandable. In most instances, this will mean use of percentile ranks and/or standard scores.

Content-Referenced and Criterion-Referenced Score Interpretation

Because content-referenced and criterion-referenced scores have only recently come into wide use (especially content-referenced scores), the Standards is relatively silent regarding their interpretation and use. However, the basic principles discussed in the previous section clearly apply: The manual should present the necessary information on the development of the scores and users should study this material so that they can interpret the scores in an accurate and understandable manner.

Content-referenced scores, you will remember, interpret performance with reference to a specified domain of tasks, skills, and/or items. Thus, the essential element in content-referenced interpretation is a description of the domain from which the test items were drawn.

J6. Any content-referenced interpretation should clearly indicate the domain to which one can generalize. (pp. 71-72)

B3.1. In the case of tests developed for content-referenced interpretation, special attention should be given to defining the content domain in operational terms. In the case of a mastery test, the test developer's rationale for any cutting score that he suggests should be specified, or the procedures that the user might employ to establish mastery levels should be described. (p. 15)

These standards are easier to state than to apply. While a general definition of the domain usually is easy to provide - e.g., reading ability, arithmetic problem solving - when one gets more specific a myriad of questions are encountered: What do we mean by reading ability? Speed and/or comprehension? What type of reading material? Under what conditions will the sample of reading be obtained? And so on. As an examinee's "reading ability" may vary under different conditions, the domain must be clearly specified. The problem is to specify the domain in enough detail so as not to be ambiguous, yet not so narrowly as to prevent generalizations beyond the particular sample of material tested.

As with norm-referenced scores, we also need a method of expressing performance. Two approaches can be used. One is to express performance on a continuum, say, as a percentage score. The other is to use some predetermined standard of mastery - a cutting score that indicates whether or not the person can perform the skill at a defined minimal level of proficiency. For example, a student might be required to solve 8 of 10 quadratic equations correctly, write a 200-word descriptive passage with no more than three grammatical errors, or type 150 words within 5 minutes with no more than one error. How this mastery score is determined is beyond the scope of this discussion. Suffice it to say that it is a matter of intense current controversy.[4]

In criterion-referenced score interpretation, performance is expressed in terms of performance on an outside (usually nontest) variable. The most common way of doing this, though by no means the only way, is through use of expectancy tables.

D3.1. In the case of tests used for prediction, expectancy tables or experience tables translating obtained scores into probabilities of success or into proficiency levels should be included whenever possible. (p. 22)

J5.4. Local normative data or expectancy tables should

ordinarily be developed, if possible, when administrative decisions are based on test scores. (p. 71)

The process of deriving criterion-referenced test scores consists of obtaining test scores, obtaining criterion scores, determining the relationship between test and criterion scores, then expressing the scores in such a manner (through expectancy tables or other means) that the test score can be interpreted in terms of expected criterion performance. In this situation, as with other methods of interpreting scores, a measure of error (here errors in prediction) should be included whenever feasible.

Note that criterion-referenced scores are most appropriate with predictive or decision-making uses of tests. In these situations, there are other important considerations besides just the accuracy of the test scores. We will discuss some of these in Chapter 5.

The Test Interpretation Process

In the previous sections of this chapter we have discussed the information needed for accurate interpretation of test scores. In this section we will consider how to communicate this information to the client[5] in a correct and understandable manner. Clear presentation of scores, as you will see, requires both knowledge of tests and communication skills. Unless the test interpreter can communicate the test results to clients so that they can understand them and their implications, all our previous work will be for naught.

Who should receive scores? We have previously discussed the qualifications of test users, be they test developers, test interpreters, or people who evaluate tests (see G1, G2, G3). So, too, we must be concerned with the qualifications of the people to whom the test scores are reported.

J2. Test scores should ordinarily be reported only to people who are qualified to interpret them. If scores are reported, they should be accompanied by explanations sufficient for the recipient to interpret them correctly. (p. 68)

Note, particularly, two aspects of this standard. First, the test user has an obligation to determine who is qualified to receive information about test scores. Second, even if a recipient is deemed qualified, the user must not just present the scores, he must explain and interpret them.

J2.2. A system of reporting test results should provide interpretations. (p. 69)

This means, in some situations, we may communicate the meaning or implications of the test scores without reporting exact scores. For example, we might tell a student that she ranks in the top 10% of high school juniors on an ability test or that on the basis of her test scores she qualifies for an accelerated section of a course, yet not report the exact test scores to her. What we have done, instead, is to tell her the meaning of the score and its implications.

Who are qualified recipients? This will, of course, vary with the situation and purpose of testing. Certainly, test takers have a right to know their scores and their interpretation.

> J2.1. An individual tested (or his agent or guardian) has the right to know his score and the interpretations made. In some instances, even scores on individual items should be made known. (p. 69)

As the Standards points out, strictly speaking this is an ethical standard rather than a standard of competent test use. Furthermore, implementation of this standard sometimes may be in conflict with good test use; e.g., releasing data on test items may harm test security.[6] In most cases, however, letting the client know the test score and its interpretation will suffice.

If a test is used for decision-making purposes, the person(s) responsible for making the decision will need to know the scores and/or their interpretation. If they do not have the background knowledge necessary to understand the implications of the scores, the test user will have to provide training or phrase interpretations in a manner that requires little technical expertise for understanding.

At the other extreme, people should not have access to scores just to satisfy their curiosity. It is the in-between situations that are unclear. For example, does a particular classroom teacher really need to know his students' scores on an ability test? Does the school principal? Who should be told students' scores on final exams? In all cases, we must balance the individual's rights against the need of the institution for information necessary to improve its operation. One solution, when in doubt, is to ask the individuals (or their parents) to give their informed consent to release the test scores to qualified users.[7]

Some tests are designed to be "self-interpreting"; that is, test results are reported directly to the test taker without an intermediary interpreting the scores. Although the

Standards warns against implying that tests can be "self-interpreted" (B4.1), the sense of the standard is that scores should not be reported without accompanying aids for interpreting scores. That is, test takers should receive sufficient information so that they can interpret their own scores.[8]

A related situation is computer-based test interpretations. Some companies provide test score interpretations based (presumably) on actuarial data. The users of these systems, should be familiar with their development, particularly the data base from which the interpretations are derived.

> J4. A test user should examine carefully the rationale and validity of computer-based test interpretations. (p. 69)

This information should be available from the provider of the service.

> B6. Test developers or others offering computer services for test interpretation should provide a manual reporting the rationale and evidence in support of computer-based interpretations of scores. (p. 17)

If such systems are used, the test user should ensure that the interpretations meet all the standards required for appropriate test interpretation.

Guidelines for test score interpretation. There are three overriding guidelines to consider when interpreting any test score. The first is that a test score is only a "best estimate" of a person's performance; an estimate obtained in a particular situation under a particular set of conditions.

> J1. A test score should be interpreted as an estimate of performance under a given set of circumstances. It should not be interpreted as some absolute characteristic of the examinee or as something permanent or generalizable to all other circumstances. (p. 68)

This standard has several important implications. One is that a score cannot be considered in isolation.

> J1.1. A test user should consider the total context of testing in interpreting an obtained score before making any decisions (including the decision to accept the score). (p. 68)

This is clearly explained by the comment following this standard:

> The standard is that one must avoid the abdication of responsibility by relying exclusively on an obtained score. Users should, in particular, look for contaminating or irrelevant variables that may have influenced obtained test scores; for example, in testing to classify school children, scores may be influenced by behavior problems, visual or hearing defects, language problems, and racial or cultural factors, as well as by ability. (p. 68)

That is, the user must look at the circumstances and conditions under which the score was obtained, not just the score itself.

Scores are also only best estimates because there are various sources of errors in scores. Some of these arise from measurement errors (unreliability), others because tests are never perfectly valid. Thus, scores are better interpreted as ranges or bands rather than as exact points.

> J2.2.1. Scores should ordinarily be interpreted in light of their confidence intervals rather than as specific values alone. (p. 69)

> D5.4. Normative data should be provided in a form that emphasizes the fallibility of an obtained score. (p. 23)

Most test publishers, recognizing the effects of unreliability, provide data on the standard error of measurement in the manual and/or incorporate this information in the test reporting process. For example, some achievement and ability tests report scores as ranges or bands rather than as exact points. This is a desirable procedure.

> D5.2.2. The choice of a standard scale should be based on either the standard error of measurement or raw scores or on some other basis that is clearly defined. (p. 23)

> F1. The test manual or research report should present evidence of reliability, including estimates of the standard error of measurement, that permits the reader to judge whether scores are sufficiently dependable for the intended uses of the test. If any of the necessary evidence has not been collected, the absence of such information should be noted. (p. 50)

If this information is not reported in the manual, the test user should still communicate the impression that the scores are only best estimates, not exact points on a continuum. Less common is the recognition that invalidity may be a cause

of error (E8). Either case, however, leads to the conclusion that test scores are only best estimates.

In other situations, scores are best estimates because they are interpreted with reference to a less than completely appropriate norm group or when no appropriate norm group is available. Another case is when a new or revised form of a test is devised, yet interpretations are based on data collected from a previous form (see D6). For example, a large body of research data has been collected on the WISC; rather than ignoring this research, a user may want to consider it when interpreting scores from the WISC-R. However, he should realize that the data may not be completely applicable to the WISC-R. A third case occurs when we equate scores on different tests. For example, some colleges require applicants to take the ACT and others require the SAT. Rather than asking a prospective student to take both tests, performance on one test is sometimes estimated from scores on the other. When doing so, even if tables of comparable scores are available, the user should realize that the interpretations will include not only the measurement errors in each score, but also those inherent in the equating process.

Second, an obtained score tells how well the person performed, not why they performed as they did. To the extent that the test is reliable and valid, the obtained score gives an accurate picture of the test taker's performance at a particular time under a particular set of circumstances. However, with rare exceptions, the score does not tell why the person obtained the score he did. This is an important consideration in many educational testing situations. For example, if Mary obtains a relatively low score on a spelling test, we have an estimate of her current spelling ability. The score, however, does not tell us why she spells poorly. True, looking at individual items may give some clues. But the test score does not provide any direct evidence of the reasons for the performance.

When dealing with any particular score, the test user should be aware that there may be various ways to interpret that score.

> J7. The test user should consider alternative interpretations of a given score. (p. 72)

This means that the interpreter should always consider the reasons for the score, not just the score. For example, a score on an achievement test might be a function of test anxiety, deliberate faking, prior knowledge of the test items, inadequate understanding of test directions, test wiseness,

or of numerous other conditions.

Third, a test score is only one of a wide variety of sources of information. Thus, decisions and/or interpretations should never be made on the basis of test scores alone; we should also consider other measures of the same characteristic and measures of other relevant characteristics. For example, decisions to place students in different sections of courses should not be made solely on the basis of intelligence or achievement test scores; when making these decisions, users should consider other sources of information such as the students' backgrounds, performance in prerequisite courses, interests, and goals.

> J8. The test user should be able to interpret test performance relative to other measures. (p. 73)

Although this standard was written with reference to interpreting scores in terms of criterion data, it applies equally as well to situations involving use of other sources of information about the test taker.

Communicating information about test scores. Given these guidelines, you should now be ready to interpret test scores to the client. Here the suggestions of Brown (1976, pp. 219-220) are appropriate.

- Use language the client understands. Just because you understand a term (e.g., standard score) doesn't mean the client will. Thus, you will have to explain the scores in nontechnical language. When in doubt, ask the client whether he understands your interpretation. Better yet, ask him to tell you what your interpretation meant.

- Be sure the client knows what the test measures or predicts. Here, again, we do not need a detailed technical explanation, just the important implications. Explain scale names or labels which may be misinterpreted; e.g., ones like dominance, masculinity-femininity, or intelligence.

- If scores are norm-referenced, be sure the client knows what groups are being used as a basis for comparisons. The reason for this guideline is, of course, that scores will have different meanings when interpreted with reference to varying groups.

- Be sure that the client recognizes that scores are only "best estimates" of performance.

- Be sure that the client knows how the scores will be

used. This is particularly important when scores will be used in placement and selection decisions.

- Consider what impact knowing the scores will have on the client. While we generally recognize that lower-than-expected scores may threaten a client, scores that are higher than expected may also affect the client. Furthermore, test interpreters must not only recognize the possibility of adverse reactions to the test scores, they must be prepared to handle these reactions.

- Implicit in many of the previous guidelines is a final point: Let clients be active participants in the test interpretation process. After all, the scores are theirs, not yours, and the decisions to be made will affect their lives more than yours. Thus, at all stages of the (test interpretation) process you should solicit their reactions, encourage them to ask questions, and help them to reflect on the implications of their scores. Although a test score is only one limited piece of information, finding out the score may set off a chain of events that drastically changes an individual's life. Thus, you must be assured that they thoroughly understand the meanings and implications of their scores. Unless clients are active participants in the test interpretation process, you may not be aware of how well they understand their scores.

Notes

1. As will be discussed later, a given test score can be compared to more than one norm group. For the present we will only discuss comparisons to a single norm group. The points discussed, however, apply to any and all norm groups used as a basis for score interpretations.

2. For a discussion of local norms see, e.g., Brown (1976, chap. 10), Ricks (1971), or Thorndike & Hagen (1977, chap. 4).

3. For discussions of types of norm-referenced scores see, e.g., Anastasi (1976, chap. 4), Angoff (1971), Brown (1976, chap. 10), Cronbach (1970, chap. 4), Ebel (1972, chap. 19), Mehrens & Lehmann (1978, chap. 6), Seashore (1955), or Thorndike & Hagen (1977, chap. 4).

4. See Ebel (1972, chap. 19), Millman (1973, 1974) and Hambleton et al. (1978).

5. In the following discussion we will use the term "client" to refer to any person who receives the test score

interpretation, whether they are the test taker or some other person.

6. Legislation, such as recently passed in New York, requires disclosure of individual items and their answers. It is still too early to tell what implication this law will have for the administration and use of secure tests.

7. Federal legislation also puts restrictions on the release of personal information, including test scores. See Privacy rights of parents and students (Federal Register, 17 June 1976, 24662-24675).

8. Most college, graduate, and professional school testing programs provide interpretive aids to test takers with their score reports. For example, one college admissions testing program provides a 48 page explanation with the test score report. They also suggest other resources which are available to aid in interpreting scores (e.g., college counseling services).

V. PROBLEMS AND ISSUES IN TEST USE

In the previous chapters we have discussed various aspects of the assessment process: the development and selection of assessment instruments, administration and scoring procedures, reliability and validity, and the interpretation of test scores. In this chapter we will consider how the standards relate to the use of tests in assessment programs. We will also look again at the problem of test fairness and discuss some ethical issues involved in assessment.

To set the context for our discussion, we will briefly review some major points of the previous discussion. In Chapter 1 we differentiated between two types of test use, decision-making uses and descriptive uses. Consider first decision-making uses. Sometimes the decision is which applicants to select (or reject) for a particular job, school, or educational opportunity. Here test scores are used, either explicitly or implicitly, to estimate (i.e., predict) performance on some variable external to the test (the criterion). The goal is to select the best qualified applicants for the job or educational program. In other cases the decision involves placement; for example, assigning students to various instructional groups on the basis of achievement or ability test scores. In this case, the underlying assumption is that people with a certain set of characteristics will perform a task or achieve a goal more effectively with one form of treatment than with another. A third use is for certification. Here we are interested in whether the individual has met some standard of proficiency that implies he is qualified to perform some task. Examples include driver's license exams, examinations required before practicing a profession, and testing to certify a child as being eligible for a special education program. A fourth situation is when test scores are used as a dependent, or criterion, measure. This use is more common when evaluating programs that when evaluating individuals. For example, achievement test scores of students may be used as a criterion measure when evaluating an instructional method or a teacher. The common thread in all cases is that test scores are used as a (at least partial) basis for making a specific decision about an individual or program.

In descriptive uses we are interested in obtaining

information that can help the test user understand or help an individual test taker.[1] For example, a teacher may use test information to plan educational programs for a student or a counselor may use test information to help a client make vocational plans. These are diagnostic and clinical uses of tests. As descriptive uses generally are most appropriate when there is a continuing relation between the test taker and test user, the information obtained from the assessment can be used at various times, for various purposes, and can be modified in the light of further information.

Because descriptive uses of tests are more open-ended and individualized than decision-making uses, there is some confusion as to how rigidly the standards we have discussed should apply to descriptive uses. The Standards comments on this problem:

> The standards of test use may not have to be so rigidly followed when the purpose of testing is the understanding of an individual. Sometimes such testing is less standardized than is usually recommended. For example, a school counselor may be interested in assessing the maximal performance capability of a single student. To get a full understanding of that student, he must be able to elicit new information, perhaps even through an embellishment of a standardized test, to seek the broadest possible understanding of the level of mastery and of the generalizability of the situations in which mastery can be demonstrated. Interpretation of test scores in such cases is not made in terms of norms but in terms of the counselor's analysis of what mastery of a particular skill entails . . . In short, exploration of an individual case is different from standardized testing. The user who develops test embellishments must know the difference; that is, he must have a clear rationale for what he is doing when he departs from standard procedure, and he must be able to apply that rationale consistently and sensibly. Such individualized test use does not require less skill than does testing broadly for institutional decisions; it requires a different kind of skill. (p. 57)

In other words, descriptive uses may require other than standard testing procedures. However, when test users use other than standard procedures, they should have a consistent and logical rationale. This does not mean that the test user can ignore the recommended standards; only that the standards will be interpreted in light of the intended use of the test.

Regardless of the purpose of testing, the goal always is

to use tests, and other assessment methods, in the most effective manner. This, in turn, requires competent test users.

> The test user, in selecting, administering, scoring or interpreting a test should know his purposes, what he is doing to achieve those purposes, and their probable consequences. It is not enough to have benign purposes; the user must know the procedures necessary to maximize effectiveness and minimize unfairness in test use. He must evaluate the many factors that may have influenced test performance in light of his purposes. Where he finds that certain factors would unfairly influence performance, his procedures for using the test and interpreting the scores should be designed to minimize such influences.
>
> Competence in test use is a combination of knowledge of psychometric principles, knowledge of the problem situation in which the testing is to be done, technical skill, and some wisdom. Although it is not appropriate to tell a test user that he needs particular levels of validity and reliability, it is appropriate to ask him to ascertain that his procedures result in reasonably valid predictions or reliable classifications or otherwise conform to the purposes of his testing. (p. 6)

Thus, competent and effective test use involves careful consideration of why we are using particular procedures and constant evaluation of how effective our procedures are in attaining our goals.

With these general considerations in mind, we will now review the standards applying to the use of tests in applied situations. As we have discussed each of the component steps in detail in previous chapters, we will consider only the major steps and problems in this section; the reader is referred to the previous chapters for more detailed discussion of particular points. Our purpose here is to provide a broad overview of the more practical concerns in the assessment process.

Using Test Scores in Assessment Programs

Effective and ethical use of tests and other assessment procedures requires qualified and competent test users. But what is meant by a "qualified" test user? In a general sense, the Standards states:

> One cannot competently judge whether his intended use [of an assessment method] is among those that are

"legitimate" (however defined) without the technical skill and knowledge necessary to evaluate the validity of various types of inferences. (p. 58)

These skills are spelled out more specifically in several standards for test use.

G1. A test user should have a general knowledge of measurement principles and of the limitations of test interpretations. (p. 58)

G1.1. A test user should know his own qualifications and how well they match the qualifications required for the uses of specific tests. (p. 58)

G2. A test user should know and understand the literature relevant to the tests he uses and the testing problems with which he deals. (p. 58)

G3. One who has the responsibility for decisions about individuals or policies that are based on test results should have an understanding of psychological or educational measurement and of validation and other test research. (p. 59)

In short, qualified users know both the general principles of educational and psychological measurement and the instruments they intend to use.

Effective use of tests also requires specification of goals and objectives: Unless we know what we are trying to accomplish, we cannot evaluate the effectiveness of our efforts.

H1. The choice or development of tests, test batteries, or other assessment procedures should be based on clearly formulated goals and hypotheses. (p. 60)

H1.1. The test user should formulate goals clearly enough so that he can later evaluate his success in achieving them and can communicate that evaluation to other qualified persons. (p. 61)

To assure that the instruments are used appropriately, test users must then relate information on the particular method(s) they use to these goals.

H3. In choosing an existing test, a test user should relate its history of research and development to his intended use of the instrument. (p. 63)

> H5. Test scores used for selection or other administrative decisions about an individual may not be useful for individual or program evaluation and vice versa. (p. 63)

Furthermore, because decisions are often complex and are based on more than one characteristic or ability, the test user may have to use more than one assessment method.

> H2. A test user should consider more than one variable for assessment and the assessment of any given variable by more than one method. (p. 61)

In brief, test users should always consider their goals when selecting particular assessment methods for use. They should not fall into the habit of using the same tests or assessment methods for all purposes, nor should they measure only those characteristics that are easily and conveniently measured. Instead, specific instruments should be selected in terms of the goals of the assessment.

Even if an appropriate method (test) is selected, the information derived will not be valuable unless testing conditions are standardized.[2]

> I1. A test user is expected to follow carefully the standardized procedures described in the manual for administering a test. (p. 64)

> I1.2. A test user should maintain consistent conditions for testing. (p. 64)

> I3. A test user is responsible for accuracy in scoring, checking, coding, or recording test results. (p. 66)

The testing conditions should be such that all individuals have the opportunity to exhibit their best performance.

> I2. The test administrator is responsible for establishing conditions, consistent with the principle of standardization, that enable each examinee to do his best. (p. 64)

The emphasis on standardized conditions is deliberate for:

> ... when decisions are based on test scores, the decision for each individual should be based on data obtained under circumstances that are essentially alike for all. (p. 64)

Unless testing conditions are comparable for all test takers, comparisons of an individual's performance to any standard

(be it one based on the performance of other people or on a content mastery standard) cannot be unambiguously interpreted. As the interpretive information available for the test applies only to situations where the standard conditions were used, to the extent that the testing conditions vary from these conditions interpretations will be less solidly based.

Interpreting test scores. The crux of test use usually is in the interpretation of scores. When interpreting test scores, users should keep in mind that the quality of the interpretation will depend upon the evidence of the reliability and validity of the instrument. Thus:

> B5. Evidence of validity and reliability, along with other relevant research data, should be presented in support of any claims being made. (p. 16)

Furthermore, it is important not to confuse reliability and validity.

> J3. The test user should recognize that estimates of reliability do not indicate criterion-related validity. (p. 69)

If there is sufficient evidence that the test is reliable and valid for the intended use, the user can interpret scores. Here, users should constantly remember that scores are only estimates.

> J1. A test score should be interpreted as an estimate of performance under a given set of circumstances. It should not be interpreted as some absolute characteristic of the examinee or as something permanent and generalizable to all other circumstances. (p. 68)

In other words, when interpreting test scores the total testing context should be considered, including the testing situation, the test taker's motivation, and other variables which might influence the obtained scores.

> J1.1. A test user should consider the total context of testing in interpreting an obtained score before making any decisions (including the decision to accept the score). (p. 68)

Keeping these considerations in mind should help avoid many possible misinterpretations of test scores.

When interpreting test scores, the test user also should use an appropriate frame of reference. For norm-referenced

interpretations this means interpreting scores with reference to an appropriate norm group.

> J5. In norm-referenced interpretations, a test user should interpret an obtained score with reference to sets of norms appropriate for the individual tested and for the intended use. (p. 70)

For many uses, this will mean the development and use of local normative data (D1.1, D4). For content-referenced interpretations, the domain must be clearly specified.

> J6. Any content-referenced interpretation should clearly indicate the domain to which one can generalize. (pp. 71-72)

In either situation, the test user should recognize that any score may be interpreted in several ways and must be considered in relation to other variables.

> J7. The test user should consider alternative interpretations of a given score. (p. 72)

> J8. The test user should be able to interpret test performance relative to other measures. (p. 73)

These two standards emphasize the fact that test scores cannot be interpreted in isolation; the total context of the testing must be considered for accurate test score interpretation.

Decision-making rules. When tests are used for decision-making purposes, the first step is to specify the procedures and rules under which the decisions will be made. This involves specifying what action will be taken for persons obtaining various test scores. In selection situations, this action generally will be to accept persons obtaining scores above some specified cutting score and to reject persons whose scores are below this level. In placement situations, the score levels (ranges) leading to placement in each of the various possible treatments need to be specified.

The test user has the responsibility of specifying the rationale for use of the particular decision rules and, whenever feasible, supporting their use with empirical evidence.

> E2. A test user is responsible for marshalling the evidence in support of his claims of validity and reliability. The use of test scores in decision rules should be supported by evidence. (p. 32)

> I4. If specific cutting scores are to be used as a basis for decisions, a test user should have a rationale, justification, or explanation of the cutting scores adopted. (p. 66)

The impact of these standards is to place the responsibility for justifying their procedures on the test user; setting cutting scores arbitrarily, casually, or by adopting procedures used in other situations or organizations are not acceptable practices.

Setting a particular cutting score usually is a two-stage process. The first stage involves determining the basis for the decision rule. In some cases this may be simply specifying the number or percent of persons falling in each category. For example, in selecting students for an honors section of a freshman English course, we might want to select the top 10% of the freshman class. Alternatively, we might set our criteria in terms of the probability of certain outcomes. To continue the above example, the criterion might be that students would not be eligible for the honors class unless they had at least an 80% chance of obtaining an A or B grade in the course. Note that in both cases we have specified the basis for the decision rule - in the first case a given percent of the freshman class was placed in honors course; in the second case the number of people placed in the course was determined by the predicted outcome (i.e., their probable grade).

The second stage in the process is determining where the cutting score should be set. This involves collecting data for a sample of people, applying various decision rules (i.e., considering various possible cutting scores), and then deciding, on the basis of the empirical evidence, which cutting score maximizes the number of correct decisions. Note that once the optimal cutting score has been determined we have, in effect, a two-point scale; scores above the cutting line are considered as "pass", scores below the cutting line are considered as "fail".

The test users' responsibility does not end once a specific cutting line has been set.

> J7.1. Where cutting scores are established as guides for decision, the test user should retain some degree of discretion over their use. (p. 72)

This is another example of a standard that is easier to state than to implement effectively. On one hand, the intent of this standard is to avoid overly rigid use of decision rules

based on measures having imperfect validity; that is, to allow for variations in individual situations (cf. J7, J8). On the other hand, altering the rules too frequently, by allowing subjective judgments to override the empirical evidence, will reduce rather than enhance validity. Thus, exceptions to decisions rules should be used infrequently and only when there is persuasive evidence that we are dealing with an exceptional set of circumstances.

Questionable scores. When tests are used for decision-making, it is necessary to ensure that the scores used as a basis for decisions are accurate indications of the test taker's abilities and characteristics. One requirement is that the scores used be reliable and valid. In addition, in most instances where test scores are used as a basis for decisions (certainly in all selection situations), it is essential that test takers not be familiar with the specific test content before taking the test; otherwise validity may be destroyed. Thus, test users must make provision for maintaining test security.

> I5. The test user shares with the test developer or distributor a responsibility for maintaining test security. (p. 67)

If the test user has reason to believe that security has been breached, the test results should not be used when making a decision; instead, other assessment methods should be substituted.

> I5.1. Where a probable breach of security may invalidate test inferences, the test user should employ other methods of assessment; that is, he should seek a basis for more valid inferences. (p. 67)

> I5.3. The test user should avoid basing decisions on scores obtained from insecure tests. (p. 68)

The reasons for these standards should be obvious.

Breaches of test security represent one situation where retesting may be desirable. Another would be when there were procedural errors in the assessment process.

> J7.3. A procedure for reporting test results should include checks on accuracy and make provision for retesting. (p. 72)

Here we are not only concerned with the source of the error, we are also concerned with what is done to correct errors.

> Errors in procedures and test scoring occur; procedures should be available for checking. Retesting is one form of checking results. There should be some limits to a retesting provision. The number of allowable retests may be limited by the number of parallel forms available. Certain types of assessments (e.g., personal-history data) are inappropriate for "retesting." Moreover, a true score is likely to be closer to the mean than its corresponding obtained score. This fact has important implications for extremely low scores; they will tend to be increased in a retest.
>
> In general, however, opportunities for retesting should be permitted without major obstacles. The principle is that no one should be a victim, without recourse, of an adverse decision on the basis of faulty and correctable psychological assessment. Nor should such decisions be permanent; over a period of time, individuals should have a chance for reevaluation on the basis of new learning or new experience. (pp. 72-73)

In other words, when in doubt, retest.

The discussion in the previous paragraph leads to a more general principle: Whenever feasible, a person should have more than one opportunity to obtain a favorable decision.

> J7.2. A person tested should have more than one kind of opportunity to qualify for a favorable decision. (p. 72)

In some situations this may involve retesting with the same instrument. For example, many applicants for college, graduate, or professional schools take the required admissions tests more than once, in hopes of obtaining a higher score. In other situations, the test user might base decisions on characteristics other than those measured by the test. Again considering college admissions, persons with outstanding high school records or evidence of special talents or academic qualifications might be accepted regardless of their scores on a college admissions test. But use of alternative qualification methods must be done with due caution.

> In general, the most valid methods available should guide decisions; the subjective use of information not validated can reduce validity. When compelling information exists, however, it should not be ignored in individual cases. (p. 72)

The implication of this principle is that obtaining an unfavorable decision, in and by itself, is not necessarily

sufficient to mandate use of other forms of assessment in an attempt to qualify an individual. However, if there is other valid evidence of the individual's potential, it should not be overlooked.

Release of information. Another issue in test use is who should have access to the information obtained from an assessment procedure. The general principle is:

> J2. Test scores should ordinarily be reported only to people who are qualified to interpret them. If scores are reported, they should be accompanied by explanations sufficient for the recipient to interpret them correctly. (p. 68)

The two important aspects of this standard are that scores should be interpreted only to persons who are qualified to use them and that scores should be accompanied by explanations (J2.2). The question of who is qualified to interpret scores has been addressed previously. Although the form of the report may vary with different audiences (e.g., teachers, parents, students), the critical requirement in all cases is that the interpretation be clear and understandable to the recipient.

To guide operational procedures we always have to ask: Who should have access to the scores? In two cases the answer is clear. When tests are used for decision-making purposes, the person making the decision must know the scores (or at least their implications). Then, too, the test taker has a right to know his or her scores.

> J2.1. An individual tested (or his agent or guardian) has the right to know his score and the interpretations made. In some instances, even scores on individual items should be made known. (p. 69)

Note that this principle says that the test taker has a right to know both his score and how it will be used, e.g., what weight it will have in the decision-making process.

Questions arise when we consider who else should have access to scores. Certainly, at one extreme, persons who are motivated only by curiosity should not be given access to scores. But there is a wide middle ground that is not so clear. Consider the scores of a high school student on achievement tests. There are a wide variety of people who might have reasons for wanting to know their scores: classroom teachers, administrators, counselors, parents, potential employers, admissions officers at colleges the student applies

to, etc. Which, if any, of these people should be allowed to see the student's scores?

One answer is provided by federal legislation, in the so-called Buckley Amendment. This amendment gives students and their parents the right of access to all their educational records, including test scores. On the other hand, the legislation prohibits disclosure of these records, except with prior written consent of the students or their parents, to other than "... other school officials, including teachers, within the educational institution or local educational agency who have been determined by the agency or institution to have legitimate educational interests" or, under certain restrictions, to officials of another school in which the student intends to enroll (Federal Register, vol. 41, no. 118, 17 June 1976, p. 24673). So the guiding principle is, unless scores are reported to designated persons within the student's school system, there must be prior, informed written consent before scores can be released. The term "informed consent" implies that the test taker will know what scores will be released and for what purposes.

In situations involving other than official school records, the recommended procedure is for test users to obtain the informed consent of the test taker before releasing any information. As with school records, this consent would be based on the test taker's knowledge of what information will be released, to whom, and for what purposes. Needless to say, information about test scores should be treated as confidential information by all persons using the scores.[3]

One final comment should be made regarding release of information: Outdated and obsolete information should be eliminated from permanent records.

> J9. A test user should develop procedures for systematically eliminating from data files test-score information that has, because of the lapse of time, become obsolete. (p. 73)

As people change, because of learning or other experiences, information collected at an earlier time will become less valid. If this information is retained in files, it might be used, possibly doing a disservice to the test taker; thus outdated information should be systematically eliminated from files.[4] As the Standards points out, not all data are equally susceptible to obsolescence. For example, subjective information probably has a shorter useful life than more objective information, and information about young children probably becomes obsolete faster than information collected on adults.

In all cases, test users should not overestimate the useful life of any piece of information, be it test scores or other data.

The need for evaluation. Regardless of the purposes of assessment program, or the specific instruments used, test users should constantly monitor and evaluate their use of tests.

> G5. Institutional test users should establish procedures for periodic internal review of test use. (p. 60)
>
> E1.3. To insure the continued correct interpretation of scores, the validity of suggested interpretations should be rechecked periodically; test developers should report results in subsequent editions of the manual. (p. 32)
>
> E11. To the extent feasible, a test user who intends to continue employing a test over a long period of time should develop procedures for gathering data for continued research. (p. 45)

The reasons for these recommendations, though hopefully obvious, are clearly explained by comments accompanying these standards.

> ... the objectives of academic programs change. The difficulty and psychological meaning of test items will also change. (p. 32)
>
> The competent use of tests and test scores requires regular review of procedures and of concepts that may change with the advent of new knowledge. A practice that might have been considered acceptable or appropriate at an early period may be found to be either harmful or ineffective in light of subsequent findings in psychometric theory or criticisms of test use. The review should examine the soundness of procedures used in test administration, the modes of assessment, the bases for inferences drawn from test scores, and the relative quality of the various validation strategies. (p. 60)

Note that this comment applies to all aspects of the assessment procedure, not merely to the need for collecting validation data on the inferences drawn from test scores.

Test Bias

One of the most complex and confusing issues in educational and psychological assessment is test bias. Although

most discussions have focused on sex and race biases, there are other possible sources of bias in the assessment process. For example, certain types of testing conditions may be unfair to persons having particular physical limitations or characteristics. Thus, test users must be constantly alert to possible sources of bias in their procedures.

> G4. Test users should seek to avoid bias in test selection, administration, and interpretation; they should try to avoid even the appearance of discriminatory practice. (p. 60)

As the Standards points out, this is a difficult standard to apply. One reason is that there are various definitions of what constitutes bias. Furthermore, we do not fully understand the factors that cause biases to occur, nor do we have the necessary psychometric techniques (at least agreed upon ones) to identify, control, and correct for biases. Finally, and by no means least important, any discussion of test bias will involve legal and ethical issues, as well as psychometric ones. Yet the possibility of test bias is a problem that must be dealt with. The competent test user must keep abreast of developments in this area and be sensitive to the problems and to the feelings of examinees.

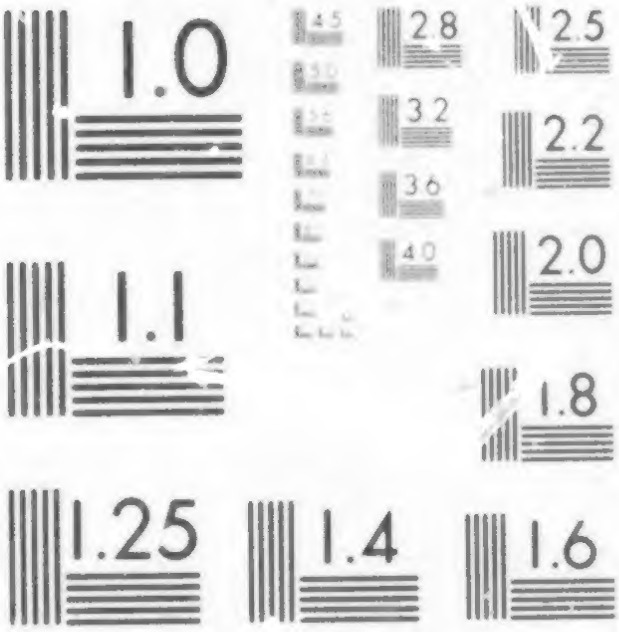
Types of test bias. Although there is no universal definition of what constitutes test bias, a distinction made by Reschley (1978, 1979) helps clarify the issues. Reschley described four types of possible test bias: content bias, atmosphere bias, bias in use: prediction, and bias in use: social consequences. We will consider each type and see what recommendations the Standards makes about each.

Content bias refers to the situation where members of various subgroups have not had equal opportunity to learn the material covered by the test. As interpretations of maximal performance tests are based on the assumption that all test takers have had the opportunity to learn the material tested, to the extent that members of any subgroup[5] have not had the opportunity to learn the material, the test may be unfair. Several parts of the Standards address this question.

> E9. A test user should investigate the possibility of bias in tests or in test items. ... (p. 43)

> E12.1.2. Test content should be examined for possible bias. (p. 45)

> B1.3. The manual should call attention to marked influences on test scores known to be associated with region,



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

> socioeconomic status, race, creed, color, national origin, or sex. (p. 14)
>
> Bias may exist where items do not represent comparable tasks and therefore do not sample a common performance domain for the various subgroups. (p. 45)

These comments suggest that whenever there is evidence that various subgroups perform differently on items, the possibility of bias should be investigated.

One could ask whether evidence of group differences is sufficient to judge a test to be biased. Some people would argue that evidence of differential performance per se is sufficient to establish bias; others would argue that group differences alone are not sufficient, we must also consider why the differences occurred. Here the questions of opportunity to learn becomes crucial. If all test takers have not had the opportunity to learn the material tested, then the test will be unfair to those who have not had the opportunity to learn. However, if all test takers have had sufficient opportunities to learn, yet the performance of various subgroups was different, we would not have persuasive evidence of bias in test content.

The situation is further complicated by the fact that a content-biased test may be a valid and useful measure in a particular situation. For example, Spanish-speaking students have not had the same opportunity to learn to read English as native-born students; yet a test of knowledge of road signs written in English would be an appropriate part of a driver's license examination. Here, use of the test is necessary, even though it places one group (those who do not read English well) at a disadvantage. This example also illustrates how content bias is intertwined with test use. That is, what is defined as content bias may vary depending upon how the test is to be used.

The practical issue is how to detect and eliminate content biases. One way is to have qualified members of the groups who might be adversely effected examine the items for evidence of bias. In order to maintain test security (if this is important) these should be people other than the ones to be assessed. If these qualified judges find some items objectionable, they can be eliminated and other items substituted. Another approach is to examine item statistics for evidence of bias; e.g., do the items have the same relative order of difficulty in various subgroups? This approach has the advantage of being based on empirical evidence, not just impressions. In either case, test developers and users must be alert to

possible biases in items.

Reschley's second type of bias, atmosphere bias, refers to the effects of the testing conditions on the test taker's performance. Particular examples would be the type of motivation elicited and factors related to the examinee-examiner interaction. However, one should also be alert to possible biases in evaluating and scoring responses, particularly when scoring free response items such as on individual intelligence tests. In general, the goal is to eliminate tester effects.

> H4. In general a test user should try to choose or to develop an assessment technique in which "tester-effect" is minimized, or in which reliability of assessment across testers can be assured. (p. 63)

The point is: The less the influence of the tester on scores, the less chance for atmosphere bias. These tester effects can be minimized by using standard testing conditions.

Here again, however, things are not as simple as they might seem on first blush. For example, directions on maximal performance tests stress doing the best one can; that is, they assume test takers will be motivated to exhibit their maximal performance. Some people would argue that while this assumption may hold true for middle-class children, it will not necessarily be true for children from other subcultures or socioeconomic levels. In addition, the nature of the examiner-examinee interaction may have significant effects on the level of the performance exhibited. This would be particularly true on tests where there is a large amount of interaction, such as on individually administered intelligence tests. To counteract possible biases, some people suggest that examiners should always be selected from the same subgroup (e.g., race, sex, or ethnic group) as the person tested. While evidence on the magnitude of tester effects is contradictory, test users should be alert to the possibility of such effects and strive to attain testing conditions consistent with the purposes of the assessment.

The third category, bias in use: prediction, refers to the accuracy of predictions made from the test. From this viewpoint, a test is biased if the predictions or decisions based on the test scores vary for different groups. Even this statement is open to various interpretations. As a consequence, a variety of statistical techniques for identifying predictive bias have been proposed.

As a practical matter, whenever feasible, test users should investigate the validity of the test within each

subgroup.

> E9. A test user should investigate the possibility of bias in tests or in test items. Whenever possible, there should be an investigation of possible differences in criterion-related validity for ethnic, sex, or other subsamples that can be identified when the test is given. The manual or research report should give the results for each subsample separately or report that no differences were found. (p. 43)
>
> B1.3. The manual should call attention to marked influences on test scores known to be associated with region, socioeconomic status, race, creed, color, national origin, or sex. (p. 14)

Such studies may have various outcomes. If there are no significant differences between groups, the same decision-making rules can be used for everyone, regardless of their group membership. If the test is valid for each group, but the regressions equations vary, separate prediction equations and/or cutting scores should be used for each group. If the procedure is not valid for one group, other assessment methods should be devised for that group. The important point is that users be alert to differences in validity for various subgroups within a heterogenous population and vary their decision-making procedures accordingly.[6]

Some critics of testing say that the previous arguments miss the point. For example, a test may be valid (e.g., for predicting school success) yet be unfair because the criterion used in the validation study was also biased. Thus, the issue becomes broader - it encompasses the uses to which the test results are put. Reschley explains what he means by bias in use: social consequences in the following statement:

> The ultimate criteria that should guide our evaluations of test bias are the implications and outcomes of test use for individuals. Succinctly stated, test use is fair if the results are more effective interventions leading to improved competencies and expanded opportunities for individuals. Test use is unfair if opportunities are diminished or if individuals are exposed to ineffective interventions as a result of tests. (Reschley, 1978, p. 33)

In Reschley's view, using test results to assign students to less than optimal programs (regardless of the fairness of test content, procedures, or validity) is an unfair use of tests. For example, using tests as a basis for assigning students to

a class for slow learners that will probably allow them less opportunity to learn than in a regular classroom would be an unfair use of a test. Note that this definition encompasses factors in addition to the test and its use, specifically the quality of the educational services provided. However the implication is clear: Fair uses of tests are those which foster the development of individuals; unfair ones hinder individual development.

A caveat. We have cautioned users to be constantly alert to possible biases in tests and test use. Such cautions must be taken seriously - both as a matter of proper test use and on ethical and legal grounds. Yet, as the Standards comments, and as we have said previously, it is equally important to avoid seeing bias when none is present. In an area so complex and confused as the area of test bias, making such discriminations is no easy task. Only by keeping abreast of current developments and constantly evaluating procedures and programs can tests be used fairly.

Ethical Issues

In our previous discussion, we have discussed user qualifications and competencies and the standards for test use. However, as the interpretations and decisions made will effect the lives of individuals, these standards also can, and must, be viewed from an ethical viewpoint.

While realizing that all readers are not psychologists, the code of ethics of the American Psychological Association (Ethical Standards of Psychologists, 1977) has much of value to say to test users. Therefore, in this section, we will quote the most relevant sections of this ethical code. Our purpose is not to imply that all test users should adhere to these ethical standards (although this would not be a bad idea); rather it is to present some of the ethical issues that test users must confront when using tests.[7] We will present these ethical standards without comment as their meaning should be clear after reading the previous sections of this booklet.

> Principle 1. Responsibility. In their commitment to the understanding of human behavior, psychologists value objectivity and integrity, and in providing services they maintain the highest standards of their profession. They accept responsibility for the consequences of their work and make every effort to ensure that their services are used appropriately.
>
> Principle 2. Competence. The maintenance of high

standards of professional competence is a responsibility shared by all psychologists in the interest of the public and the profession as a whole. Psychologists recognize the boundaries of their competence and the limitations of their techniques and only provide services, use techniques, or other opinions as professionals that meet recognized standards. Psychologists maintain knowledge of current scientific and professional information related to the services they render.

2d. Psychologists with the responsibility for decisions involving individuals or policies based on test results have an understanding of psychological or educational measurement, validation problems and other test research.

Principle 5. Confidentiality. Safeguarding information about an individual that has been obtained by the psychologist in the course of his teaching, practice, or investigation is a primary obligation of the psychologist. Such information is not communicated to others unless certain important conditions are met.

5b. Information obtained in clinical or counseling relationships, or evaluative data concerning children, students, employees, and others are discussed only for professional purposes and only with persons clearly concerned with the case. Written and oral reports should present only data germane to the purposes of the evaluation and every effort should be made to avoid undue invasion of privacy.

5f. The psychologist makes provision for the maintenance of confidentiality in the preservation and ultimate disposition of confidential records.

Principle 6. Welfare of the consumer. Psychologists respect the integrity and protect the welfare of the people and groups with whom they work. When there is a conflict of interest between the client and the psychologist's employing institution, psychologists clarify the nature and direction of their loyalties and responsibilities and keep all parties informed of their commitments. Psychologists fully inform consumers as to the purpose and nature of an evaluative, treatment, educational or training procedure, and they freely acknowledge that clients, students, or participants in research have freedom of choice with regard to participation.

Principle 8. Utilization of assessment techniques. In the development, publication, and utilization of

psychological assessment techniques, psychologists observe relevant APA standards. Persons examined have the right to know the results, the interpretations made, and, where appropriate, the original data on which final judgments were based. Test users avoid imparting unnecessary information which would compromise test security, but they provide requested information that explains the basis for decisions that may adversely affect that person or that person's dependents.

8a. The client has the right to have and the psychologist has the responsibility to provide explanations of the nature and purposes of the test and test results in language that the client can understand, unless, as in some employment or school settings, there is an explicit exception to the rights agreed upon in advance. When the explanations are to be provided by others, the psychologist establishes procedures for providing adequate explanations.

8b. When a test is published or otherwise made available for operational use, it is accompanied by a manual (or other published or readily available information) that fully describes the development of the test, the rationale, and evidence of validity and reliability. The test manual explicitly states the purposes and applications for which the test is recommended and identifies special qualifications required to administer the test and interpret it properly. Test manuals provide complete information regarding the characteristics of the normative population.

8c. In reporting test results, psychologists indicate any reservations regarding validity or reliability resulting from test circumstances or inappropriateness of the test norms for the person tested. Psychologists strive to ensure that the test results and their interpretations are not misused by others.

8d. Psychologists accept responsibility for removing from clients' files test score information that has become obsolete, lest such information be misused or misconstrued to the disadvantage of the person tested.

8e. Psychologists offering test scoring and interpretation services are able to demonstrate that the validity of the programs and procedures used in arriving at interpretations are based on appropriate evidence. The public offering of an automated test interpretation service is considered as a professional-to-professional

consultation. The psychologist makes every effort to avoid misuse of test reports.

The thrust of these ethical principles is clear: Psychologists (or other test users) must know their own competencies and qualifications, use valid methods in an appropriate manner, and always be cognizant of the rights of and their responsibilities to the people with whom they are working.

A Final Word

In this booklet we have discussed the Standards for educational and psychological tests, with particular emphasis on the use of tests (and other assessment methods) in educational settings. While we have discussed many specific problems and issues, the major import of the discussion can be summarized by two principles. First, competent, effective, and ethical use of tests requires a qualified person, using valid methods in an appropriate manner, to attain specified goals. Second, the goals of any assessment procedure should be to preserve the dignity of individuals and foster the development of their potentials.

Notes

1. Test takers can, of course, make similar analyses of their own scores. Our discussion, however, will be oriented to test users.

2. The primary test user has the responsibility for training and supervising the work of clerks, semi-professionals, or other persons who are involved in any aspect of an assessment program (G3.1). He also has responsibility to provide all necessary information about the testing program and procedures to other persons concerned with the testing program (I2.1).

3. For further discussions see, e.g., Anastasi (1976, pp. 49-55), Privacy and behavioral research (1967), Ruebhausen & Brim (1966), and Guidelines for the collection, maintenance, and dissemination of pupil records (1970).

4. If information is retained for research purposes, it should, whenever possible, be in such a form that identification of specific individuals is not possible.

5. Similar argument to those given in this section could also be made for individuals (Novick & Ellis, 1977). However, we will discuss the issues as they apply to subgroups within a broader population. We will also only consider maximal performance tests.

6. For further discussions of test bias, especially of the psychometric problems, see, e.g., Cleary et al. (1975), Cole (1973), Darlington (1971), ETS Proceedings (1973, 1978), Hunter & Schmidt (1976), Lerner (1978), Lieberman & Weiner (1975), Petersen & Novick (1976), Schmeieser & Ferguson (1978) and Thorndike (1971b).

7. See also Standards for providers of psychological services (1977).

GLOSSARY

ACT. The American College Testing Program. As used herein, refers to the college admissions test portion of the ACT Assessment Program.

Achievement test. A test measuring the extent to which a person has acquired certain information or mastered certain skills, usually as a result of specific instruction.

AERA. The American Educational Research Association. A group of educators, psychologists, and other professionals interested in research on educational problems.

Alternative-choice item. An item on which the test taker's task is to select the correct choice from a specified list of response alternatives. Examples include multiple-choice, true-false, and matching items.

APA. The American Psychological Association. A national professional organization for psychologists.

Assessment. A systematic basis for making inferences about characteristics of people, programs, etc. Common assessment methods include tests, rating scales, personality inventories, and structured observations.

Atmosphere bias. Occurs when the testing conditions (e.g., the examiner-examinee interaction) are such that certain examinees are treated differently than others, thus do not have a fair chance to exhibit their abilities.

Bias in use. Includes two types of test bias. Prediction bias occurs when the test does not predict equally (i.e., is not equally valid) for all subgroups within a broader population. Bias re social consequences occurs when certain subgroups are denied access to the optimal treatment or opportunities as a result of the manner in which test scores are used to make decisions.

Coefficient of equivalence. See parallel forms reliability.

Coefficient of stability. See test-retest reliability.

Coefficient of stability and equivalence. A measure of reliability obtained by correlating scores on two forms of a test whose administrations are separated by a period of time.

Composite score. A score obtained by combining scores from two or more measures, subtests, scales, or samples of items.

Concurrent validity. A measure of the relationship (usually a correlation coefficient) between two variables, a test (predictor) and criterion, measured at the same point in time.

Content bias. Occurs when members of various subgroups within a broader population have not had equal opportunities to learn the material covered on the test.

Construct validity. The degree to which a test measures the trait or psychological characteristic it is designed to measure.

Content validity. The degree to which the items on a test represent the underlying content and/or skill domain.

Content-referenced scores. Scores which are interpreted with reference to the degree of mastery of some defined content and/or skill domain. Examples include percent correct and mastery scores.

Correction for attenuation. A statistical correction of the correlation between two measures which takes into account the unreliability of each individual measure. Gives the hypothetical correlation between error-free measures of the two variables.

Correlation coefficient. A measure of the degree of relationship between two sets of measures obtained from the same set of individuals. Indicates the percent of variance the measures share in common and how accurately scores on one variable can be predicted from scores on the other. Unless otherwise specified, usually refers to the Pearson product-moment correlation coefficient.

Criterion. The measure or standard against which a test is evaluated. The variable the test is designed to correlate with or predict.

Criterion-referenced measurement. Measurement which compares

an individual's performance to a standard based on level of mastery of some defined content and/or skill domain. (In this booklet, called content-referenced measurement.)

Criterion-referenced scores. Scores which express performance in terms of status or level of performance on some non-test criterion measure. For example, interpreting a test score as indicating the test taker will probably obtain a B average in college.

Criterion-related validity. The extent to which test scores predict performance on some outside measure (the criterion). Most often indicated by a validity coefficient.

Cross-validation. The process of evaluating the validity of a predictor by repeating the validation process on another sample drawn from the same population as used in the original validation study. Designed to minimize the effects of error producing factors particular to a specific sample.

Cutting score. In criterion-related validity studies, the test score which divides the group into two or more categories in terms of predicted criterion performance. In selection, the score that separates those predicted to succeed from those predicted to fail. In placement decisions, the score(s) that determine what treatment condition an individual will be assigned to.

Decision-making accuracy. An index of the degree to which inferences from test scores allow for correct decisions to be made about the selection or placement of individuals. Usually expressed as the proportion or percentage of people correctly classified on the basis of their test scores.

Diagnostic test. A test whose scores indicate the test taker's relative strengths and weaknesses in the component aspects of the skill being measured.

Difference scores. Scores derived by taking the difference between scores on two variables. For example, the difference between verbal and quantitative scores on an ability test or between the scores on two administrations of the same test.

Domain. The hypothetical population of knowledge, skills, or items of interest.

ETS. Educational Testing Service. A research and test development organization.

Evaluation. Judgment of the quality, value, or worth of some performance or program. Measurements (e.g., test scores) may provide a basis for evaluation, but evaluation always includes an additional subjective judgment of the value of the performance.

Expectancy table. A table showing the proportion, or percentage, of people at each predictor score level who attain each outcome or criterion score.

Face validity. Whether the test, on casual inspection, appears to measure the ability or characteristic it is designed to measure.

Free-response item. An item on which the test taker must construct or supply the response. Examples include short answer, completion, definition, and essay items.

Grade-equivalent score. A score which reports performance in terms of the performance of an average child at a given grade level. For example, a grade-equivalent score of 4.6 on a reading test means that the child obtained the same score as the average child in the 6th month of 4th grade.

Group separation index. Any statistic describing the difference in average performance between two or more groups.

Homogeneity. The degree to which all items on a test measure the same trait or characteristic. A type of internal consistency.

Informed consent. Refers to test takers giving permission to release test scores or other information. For consent to be informed, the test taker must know what information will be released, to whom, and for what purposes.

Intelligence quotient (IQ). A measure of an individual's relative performance on a general intelligence test. IQs on the major individual intelligence tests have a mean of 100 and a standard deviation of 15 (WISC-R) or 16 (Stanford-Binet) IQ points at each age level.

Internal consistency. A type of reliability which indicates whether the test items or parts of the test are intercorrelated, thus all measure the same trait, ability, or characteristic.

Item. A single question or exercise on a test.

Item bank. A collection of test items measuring designated

knowledge, skills, or instructional objectives. Items can be sampled from this pool to form a test meeting the specifications of an individual test user.

Local norms. Norms developed to represent the distribution of performance within a particular, specified local population.

Mastery score. The level of proficiency required to say that a test taker has mastered (i.e., knows) the material in a particular domain.

Maximal performance test. A test designed to measure the test taker's maximal (i.e., best possible) level of performance. Examples include ability, achievement and aptitude tests.

Measurement. The process of assigning numbers to performance according to specified rules and operations.

Measurement error. Inconsistencies in test scores, from occasion to occasion, attributable to the effects of variables that operate in a nonsystematic (i.e., random) fashion.

NCME. The National Council on Measurement in Education. A professional organization of educators and others concerned with the application of tests to educational matters.

Normative data. A distribution of scores which describes the test performance of a specified group, such as students of a given grade or age. These data are usually obtained by testing a representative sample of the specified population.

Norm group. The sample of test takers whose scores constitute the normative data. Usually a representative sample of some specified population.

Norm-referenced measurement. Measurement resulting in scores that are interpreted by comparing the test taker's performance to that of others in a specified norm group.

Norm-referenced scores. Scores reporting performance in terms of the test taker's relative position compared to a specified norm group. Examples include percentile ranks, standard scores, and deviation IQs.

Objective (instructional). A statement of the intended outcomes of an instructional program. Behavioral objectives include specification of the knowledge or skill to be

demonstrated, the conditions under which the performance will be demonstrated, and the minimal acceptable level of proficiency.

Objective scoring. Scoring procedures which ensure high agreement between competent (i.e., trained) scorers.

Parallel (test) forms. Two or more forms of a test which measure the same content and/or skills with the same relative emphasis on each content/skill area. All forms of the test should have comparable means, variability, and distributions, as well as content. (If the latter conditions are met, often called equivalent forms.)

Parallel forms reliability. A measure of reliability derived from the correlation between two or more parallel forms of a test. Also called the coefficient of equivalence.

Percentile rank. The test taker's relative rank, in percentage terms, within a norm group. The percent of scores in a distribution equal to or lower than a particular score. For example, a percentile rank of 83 indicates that the test taker scored as high or higher than 83% of the people in the norm group.

Placement. The situation where the decision is which of several alternative treatments (e.g., jobs, classes) a person should be assigned to, with each person being assigned to a treatment. For example, assigning students to various levels within a math sequence.

Predictive validity. The relationship, usually expressed as a validity coefficient, between test (predictor) scores and scores on a criterion measure when a period of time intervenes between the collection of test and criterion scores. A variety of criterion-related validity.

Reliability. How consistently a test measures over time, occasions, and/or samples of items. The degree to which test scores are influenced by errors of measurement. Usually measured by a reliability coefficient and/or the standard error of measurement.

Reliability coefficient. The correlation between scores on various forms of a test, between scores on repeated administrations of the same test, or between items (or groups of items) within a test.

Restriction of range. The situation where, because of the characteristics of a particular sample, the variability of

scores is less than that found in a broader population of interest.

SAT. The Scholastic Aptitude Test. The aptitude portion of the College Entrance Examination Board admissions testing program.

Scale. (a) A set of items, not necessarily administered as a separate part of a test, which measures a particular trait or characteristic. (b) A dimension, having quantitative units, which reflects levels of a trait or ability.

Score. The quantitative value assigned to an individual's performance on a test, subtest, scale, or group of items. Unless otherwise designated, refers to the raw score, the score directly obtained from the test such as the number correct or total number of points earned.

Score band. The range within which a person's score falls, taking into account measurement error. Usually expressed as the obtained score plus-and-minus some multiple of the standard error of measurement.

Scoring key. A list of correct responses and acceptable alternative responses. Test takers' responses are compared to the key to determine their test scores.

Selection. The situation where there are more applicants than available positions. The goal is to select the most qualified persons (i.e., the ones with the greatest probability of success) from the applicant pool.

Selection efficiency. The degree to which a predictor correctly identifies those people who will be successful (and/or unsuccessful) in performing the criterion task.

Social desirability. A response bias where test takers respond to items, not on the basis of their true behavior or attitude, but on the basis of what is considered desirable behavior by most people. Occurs on typical performance measures.

Specimen set. A copy of the test and all supporting material (e.g., answer sheets, profiles, manuals) sold by test publishers to people wanting to examine a test.

Standard error of estimate. A measure of the amount of error in predictions. The standard deviation of the distribution of prediction errors (actual criterion scores minus predicted criterion scores).

Standard error of measurement. A measure of the amount of measurement error in test scores. Its size is a function of the reliability of the test and the variability of the scores in the distribution.

Standard deviation. A measure of the variability, or dispersion, in a set of test scores. The lower the standard deviation, the more the scores cluster around the mean; the higher the standard deviation, the more variable the scores are around the mean.

Standard score. A general term relating to a class of transformed scores all of which express performance in terms of the number of standard deviations a person's score is from the mean.

$$\text{Standard score } (z) = \frac{\text{raw score } (X) - \text{mean } (M)}{\text{standard deviation } (s)}$$

Standard scores are usually transformed to another scale by use of appropriate constants; i.e., $A + B(z)$.

Standardization. The process of developing a standard set of procedures for administering and scoring a test. Requires specification of the test content, directions for administration, and scoring rules.

Standardized test. A systematic sample of performance obtained under prescribed conditions, scored according to definite rules, and evaluated by reference to normative data. (Some definitions do not require interpretations with reference to normative data.)

Standards for educational and psychological tests. A publication of APA, AERA, and NCME which gives standards for the development and evaluation of tests and other assessment methods, interpretation of scores, and the use of tests.

Subtest. A set of items administered and scored as a separate part of a more comprehensive test.

Test. A systematic procedure for measuring a sample of behavior. A set of tasks or questions intended to elicit particular types of behavior when presented under standardized conditions and to yield scores that will have desirable psychometric properties such as high reliability and validity.

Test anxiety. Tension over performance to such an extent that it interferes with the test score being an accurate indication of the test taker's abilities or characteristics.

Test bias. Refers to situations where test procedures give a systematic advantage, or disadvantage, to members of some subgroup in comparison to other test takers. Bias can be present in the test content, administrative procedures, and/or interpretation and use of test scores.

Test manual. A booklet providing information about a test such as its development, administration, scoring, interpretation, reliability and validity. Some tests have more than one manual; e.g., a users manual and a technical manual.

Test plan. A specification of what is to be covered by the test items. Usually a table showing content categories as one dimension, cognitive skills (e.g., knowledge, understanding, application) as the other dimension, and showing the proportional emphasis to be given each content-skill category.

Test user. The person who chooses tests, interprets scores, or makes decisions based on test scores.

Test-retest reliability. A measure of consistency obtained by correlating scores on two administrations of the same test separated by a time interval. Also called the coefficient of stability.

Trait. A presumed characteristic on which people vary, such as intelligence, mechanical ability, dominance. Traits are characteristics which are relatively stable over time and situations.

Transformed scores. Scores which are expressed on a scale other than raw scores. Examples include percentile ranks, standard scores, IQs.

Typical performance test. A test designed to measure the test taker's usual or habitual behaviors, reactions, or attitudes. Examples include personality measures, interest inventories, and attitude scales.

Validity. The degree to which a test measures the characteristic it is designed to measure or predicts some outside criterion.

Validity coefficient. The correlation between scores on a test (predictor) and a criterion measure.

Validity generalization. The extent to which inferences made from test scores apply to other situations or populations.

The extent to which test scores can be interpreted similarly in various populations.

WISC-R. The Wechsler Intelligence Scale for Children, revised edition. An individual intelligence test for children aged 6-16.

REFERENCES

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D. C.: American Council on Education, 1971.

Anastasi, A. Psychological testing (4th ed.). New York: Macmillan, 1976.

Backer, T. E. A directory of information on tests. ERIC TM Report no. 62, 1977.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. Handbook on formative and summative evaluation of student learning. New York: McGraw-Hill, 1971.

Brown, F. G. Principles of educational and psychological testing (2nd ed.). New York: Holt, Rinehart & Winston, 1976.

Buros, O. K. (Ed.). The seventh mental measurements yearbook. Highland Park, N. J.: Gryphon Press, 1972.

Buros, O. K. (Ed.). Tests in print II. Highland Park, N. J.: Gryphon Press, 1974.

Buros, O. K. (Ed.). The eighth mental measurements yearbook. Highland Park, N. J.: Gryphon Press, 1978.

Cleary, T. A., Humphreys, L. G., Kendrick, S. A., & Wesman, A. Educational uses of tests with disadvantaged students. American Psychologist, 1975, 30, 15-41.

Cole, N. S. Bias in selection. Journal of Educational Measurement, 1973, 10, 237-255.

Cooley, W. W. Techniques for considering multiple measurements. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D. C.: American Council on Education, 1971.

Cronbach, L. J. Essentials of psychological testing (3rd ed.). New York: Harper & Row, 1970.

Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D. C.: American Council on Education, 1971.

Darlington, R. B. Another look at "culture fairness". Journal of educational measurement, 1971, 8, 71-82.

Dawes, R. M. Fundamentals of attitude measurement. New York: Wiley, 1972.

Ebel, R. L. Essentials of educational measurement. Englewood Cliffs, N. J.: Prentice-Hall, 1972.

Edwards, A. L. The measurement of personality traits by scales and inventories. New York: Holt, Rinehart & Winston, 1970.

Ethical standards for psychologists (1977 revision). Washington, D. C.: American Psychological Association, 1977.

Glaser, R. Instructional technology and the measurement of learning outcomes. American Psychologist, 1963, 18, 519-521.

Guidelines for the collection, maintainence, and dissemination of pupil records. New York: Russell Sage Foundation, 1970.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

Harris, C. W., Alkin, M. C., & Popham, W. J. (Eds.). Problems in criterion-referenced measurement. Los Angeles: Center for the Study of Evaluation, University of California - Los Angeles, 1974 (CSE monograph series in evaluation, no. 3).

Humphreys, L. G. The fleeting nature of the prediction of college academic success. Journal of Educational Psychology, 1968, 59, 375-380.

Hunter, J. E., & Schmidt, F. L. Critical analysis of the statistical and ethical implications of various definitions of test bias. Psychological Bulletin, 1976, 83, 1053-1071.

Jones, L. V. The nature of measurement. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D. C.: American Council on Education, 1971.

Kerlinger, F. N. Foundations of behavioral research (2nd ed.). New York: Holt, Rinehart & Winston, 1973.

Kerlinger, F. N., & Pedhazur, E. J. Multiple regression in behavioral research. New York: Holt, Rinehart & Winston, 1973.

Lerner, B. The Supreme Court and the APA, AERA, NCME Test Standards: Past references and future possibilities. American Psychologist, 1978, 33, 915-919.

Lieberman, M., & Weiner, M. (Eds.). The use of tests in personnel administration. Education and Urban Society, 8(1), 1975 (entire issue).

Manning, W. H. Discussion. In Proceedings of the 1977 ETS Invitational Conference (Educational measurement and the law). Princeton, N. J.: Educational Testing Service, 1978.

Martuza, V. R. Applying norm-referenced and criterion-referenced measurement in education. Boston: Allyn & Bacon, 1977.

Mayo, S. T. What experts think teachers ought to know about educational measurement. Journal of Educational Measurement, 1964, 1, 79-86.

Mayo, S. T. Pre-service preparation of teachers in educational measurement. Chicago: Loyola University, 1967 (Final report, Project no. 5-0807, U. S. Office of Health, Education & Welfare; Office of Education/Bureau of Research).

Meehl, P. E., & Rosen, A. Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. Psychological Bulletin, 1955, 52, 194-216.

Mehrens, W. A., & Lehmann, I. J. Measurement and evaluation in education and psychology (2nd ed.). New York: Holt, Rinehart & Winston, 1978.

Millman, J. Passing scores and test lengths for domain-referenced tests. Review of Educational Research, 1973, 43, 205-216.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), Evaluation in education. Berkeley, Ca.: McCutchan, 1974.

Novick, M. R., & Ellis, D. D., Jr. Equal opportunity in educational and employment selection. American Psychologist, 1977, 32, 306-320.

Nunnally, J. Psychometric theory (2nd ed.). New York: McGraw-Hill, 1978.

Payne, D. A. The assessment of learning: cognitive and affective. Lexington, Ma.: Heath, 1974.

Petersen, N. S., & Novick, M. R. An evaluation of some models for culture-fair selection. Journal of Educational Measurement, 1976, 13, 3-29.

Popham, W. J. (Ed.). Criterion-referenced measurement: An introduction. Englewood Cliffs, N. J.: Educational Technology Publications, 1973.

Popham, W. J. Criterion-referenced measurement. Englewood Cliffs, N. J.: Prentice-Hall, 1978.

Popham, W. J., & Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

Principles for the validation and use of personnel selection procedures. Dayton, Oh: APA Division 14, The Industrial-Organizational Psychologist, 1975.

Privacy and behavioral research. Washington, D. C.: U. S. Government Printing Office, 1967.

Privacy rights of parents and students. Federal Register, 17 June 1976, vol. 41, no. 118, 24662-24675.

Proceedings of the 1972 ETS Invitational Conference (Assessment in a plualistic society). Princeton, N. J.: Educational Testing Service, 1973.

Proceedings of the 1977 ETS Invitational Conference (Educational measurement and the law). Princeton, N. J.: Educational Testing Service, 1978.

Reschley, D. J. Non-biased assessment. Des Moines, Iowa: State of Iowa Department of Public Instruction, 1978.

Reschley, D. J. Nonbiased assessment. In G. Phye & D. J. Reschley (Eds.), School psychology: Perspectives and issues. New York: Academic Press, 1979.

Ricks, J. H., Jr. Local norms: When and why. New York: The Psychological Corporation (Test Service Bulletin no. 58), 1971.

Ruebhausen, O. M., & Brim, O. G., Jr. Privacy and behavioral research. American Psychologist, 1966, 21, 423-437.

Schmeiser, C. B., & Ferguson, R. L. Performance of black and white students on test materials containing content based on black and white cultures. Journal of Educational Measurement, 1978, 15, 193-200.

Seashore, H. G. Methods of expressing test scores. New York: The Psychological Corporation (Test Service Bulletin no. 48), 1955.

Standards for educational and psychological tests. Washington. D. C.: American Psychological Association, 1974.

Standards for educational and psychological tests and manuals. Washington, D. C.: American Psychological Association, 1966.

Standards for providers of psychological services. Washington, D. C.: American Psychological Association, 1977.

Thorndike, R. L. Reliability. In E. F. Lindquist (Ed.), Educational measurement. Washington, D. C.: American Council on Education, 1951.

Thorndike, R. L. (Ed.). Educational measurement (2nd ed.). Washington, D. C.: American Council on Education, 1971 (a).

Thorndike, R. L. Concepts of culture-fairness. Journal of Educational Measurement, 1971, 8, 63-70 (b).

Thorndike, R. L., & Hagen, E. P. Measurement and evaluation in psychology and education (4th ed.). New York: Wiley, 1977.

Wiggins, J. S. Personality and prediction: Principles of personality assessment. Reading, Ma.: Addison-Wesley, 1973.

INDEX

Administration of tests, 30-40
  directions, 31-33
  scoring, 35-38

Bias, 29, 61-62, 106-111
  types of bias, 107-111

Construct validity, 55, 65-67
Content validity, 54-55, 62-65
Content-referenced scores, 49, 74-75, 84-85
Criterion-referenced scores, 75, 85-86
Criterion-related validity, 54, 56-62

Directions, 31-33

Ethical issues, 111-114

Interpretation of test scores, 75-92, 99-102
  decision-making rules, 100-102
  guidelines for, 88-92
  qualifications of test interpreters, 73-74
  types of, 74-75

Local norms, 80-82

Manuals, 21-22, 26-28
Measurement errors, 49-50
Multiple assessment, 16-17

Normative data, 75-82
  local norms, 80-82
  norm groups, 76-80
Norm-referenced scores, 74, 75-84
  types of, 82-84

Reliability, 43-51
  definition, 42-43
  in test interpretation and use, 43-44, 49-51
  studies, 48-50
  types of, 44-47

Scores and scoring, 35-38, 71-93
  interpretation of scores, 74-92, 99-102
  questionable scores, 102-104
  release of scores, 39-40, 86-88, 104-106
  scoring, 35-38
  types of scores, 74-86

Security, 38-39
Selecting and evaluating tests, 24-30
Sources of information about tests, 25-28
Standards for educational and psychological tests, 5, 8-9, 12

Tests and testing
  availability of tests, 15-16
  definition of, 12-14
  desired characteristics, 12-15
  selection and evaluation of, 24-30
  sources of information, 25-28
  testing conditions, 30-40
  types and uses, 7-12, 96-99
Test users
  defined, 21
  qualifications of, 20-24
  responsibilities of, 23-24, 33-34, 53, 96-99, 111-114

Validity, 51-68
  defined, 42, 51-53
  implications for test users, 67-68
  studies, 52-53
  types of, 54-67